nature portfolio

https://doi.org/10.1038/s41586-024-07830-1

Supplementary information

The genomes of all lungfish inform on genome expansion and tetrapod evolution

In the format provided by the authors and unedited

Supplementary Information

The genomes of all lungfish inform on genome expansion and tetrapod evolution

Schartl, Manfred[1,2,3]*, Joost Woltering[4], Iker Irisarri[5], Kang Du[2], Susanne Kneitz[6], Martin Pippel[7,8,9], Thomas Brown[7,8,10], Paolo Franchini[4,11], Jing Li[4], Ming Li[4], Mateus Adolfi[1], Sylke Winkler[7], Josane de Freitas Sousa[12], Zhuoxin Chen[13], Sandra Jacinto[13], Evgeny Z. Kvon[13], Luis Rogério Correa de Oliveira[14], Erika Monteiro[14], Danielson Baia Amaral[14], Thorsten Burmester[15], Domitille Chalopin[16], Alexander Suh[17,18,19], Eugene Myers[7,20], Oleg Simakov[21], Igor Schneider[12,14], Axel Meyer[4]*

Affiliations
[1]Developmental Biochemistry, Biocenter, University of Würzburg, Germany.
[2]The *Xiphophorus* Genetic Stock Center, Texas State University, San Marcos, TX, USA.
[3]Research Department for Limnology, University of Innsbruck, Mondsee, Austria
[4]Department of Biology, University of Konstanz, Germany.
[5]Centre for Molecular Biodiversity Research, Leibniz Institute for the Analysis of Biodiversity Change, Museum of Nature, Hamburg, Germany
[6]Biochemistry and Cell Biology, Biocenter, University of Würzburg, Germany.
[7]Max-Planck Institute of Molecular Cell Biology and Genetics, Dresden, Germany
[8]DRESDEN-concept Genome Center (DcGC), Center for Molecular and Cellular Bioengineering, Technische Universität Dresden, Germany
[9]Present address: Department of Cell and Molecular Biology, Uppsala University, Sweden
[10]Present address: Leibniz Institute for Zoo & Wildlife Research, Berlin, Germany
[11]Present address: Department of Ecological and Biological Sciences, University of Tuscia, Viterbo, Italy
[12]Department of Biological Sciences, Louisiana State University, Baton Rouge, LA, United States
[13]Department of Developmental & Cell Biology, School of the Biological Sciences, University of California, Irvine, CA, USA
[14]Instituto de Ciências Biológicas, Universidade Federal do Pará, Belém, Brazil
[15]Institut für Zoologie, Universität Hamburg, Germany.
[16]University of Bordeaux, CNRS, IBGC, Bordeaux, France
[17]Department of Organismal Biology – Systematic Biology,Evolutionary Biology Centre, Uppsala University, Science for Life Laboratory, Uppsala, Sweden
[18]School of Biological Sciences, University of East Anglia, Norwich, UK
[19]Present address: Centre for Molecular Biodiversity Research, Leibniz Institute for the Analysis of Biodiversity Change, Bonn, Germany
[20]Center of Systems Biology Dresden, Germany
[21]Department for Neurosciences and Developmental Biology, University of Vienna, Austria

*Corresponding authors: Manfred Schartl (phch1@biozentrum.uni-wuerzburg.de) and Axel Meyer (axel.meyer@uni-konstanz.de)

Supplementary Information 1: Genome sequencing, assembly, and annotation

Genome sequencing

Extraction of ultra-long and long genomic DNA. Ultra-long genomic DNA of *L. paradoxa* and *P. annectens* was extracted from snap-frozen liver tissue samples using the Bionano Prep™ Animal Tissue DNA Isolation Soft Tissue Protocol (Bionano document number 30077). In brief, 75 mg of liver tissue was cut into 3 mm pieces and homogenized with a tissue grinder followed by a DNA stabilization step with ethanol. The homogenate pellet was then embedded in 2% agarose plugs cooled to 43°C. Plugs were treated with Proteinase K and RNase A and washed with 1X Bionano Prep Wash Buffer and 1X TE Buffer (pH 8.0). DNA was recovered with 2 µl of 0.5 U/µl Agarase enzyme per plug for 45 minutes at 43°C and further purified by drop dialysis with 1X TE Buffer.

In an alternative approach, ultra-long gDNA of both species was extracted making use of the Bionano SP tissue and tumor DNA isolation kit (Bionano document number 30339 version A). In brief, 10 mg of snap-frozen liver tissue was homogenized with the TissueRuptor II device (Qiagen). Cell debris was removed by filtration. After cell lysis, gDNA in lysis buffer was bound to Nanobind Disks in the presence of isopropanol. After washing, ultra-long molecular gDNA was eluted from nano discs in elution buffer.

Long genomic DNA of *L. paradoxa* for Pacific Bioscience HiFi sequencing was extracted with the Circulomics Nanobind Tissue Big DNA kit (part number NB-900-701-01, protocol version Nanobind Tissue Big DNA Kit Handbook v1.0 (11/19)) according to the manufacturer's instructions. In brief, 30-40 mg of liver tissue were minced to small slices on a clean and cold surface. Tissues were finally homogenized with the TissueRuptor II device (Qiagen) using its maximal settings. After complete tissue lysis, remaining cell debris was removed, and the gDNA was bound to Circulomics Nanobind disks in the presence of isopropanol. Long gDNA was eluted from the Nanobind discs in elution buffer.

The integrity of the ultralong and long gDNA was determined by pulse field gel electrophoresis using the Pippin Pulse™ device (SAGE Science). The majority of the gDNA was between 10 and 500 kb (long gDNA) and reaching more than 600 kb in length (ultra-long gDNA). All pipetting steps of ultra-long and long gDNA were done very carefully with wide-bore pipette tips. An overview of gDNA extraction protocols and employed applications is summarized in Supplementary table 9.

Supplementary table 9: gDNA extractions, fragment lengths of gDNA and sequencing method

Species	Extraction protocol	gDNA fragment length	Sequencing method
Lepidosiren paradoxa	Bionano agarose plug prep (soft tissue)	50 to > 600 kb	PacBio HiFi
	Bionano SP tissue and tumor kit	50 to > 600 kb	PacBio HiFi
	Circulomics Nanobind tissue kit	10 to 500 kb	PacBio HiFi
Protopterus annectens	Bionano agarose plug prep (soft tissue)	50 to > 500 kb	PacBio CLR
	Bionano SP tissue and tumor kit	50 to > 500 kb	Chromium genomic linked reads

Pacific Biosciences (PacBio) long-read sequencing. Because the HiFi technology was not available at the time when the genome of the African Lungfish (*P. annectens*) was sequenced, PacBio continuous long read (CLR) sequencing was applied in this species. Two long insert libraries were prepared as recommended by Pacific Bioscienes according to the 'Guidelines for preparing size-selected 20 kb SMRTbell™ templates making use of the SMRTbell express Template kit 2.0. In summary, ultra-long gDNA was sheared into 75 kb fragments with the MegaRuptor™ device (Diagenode) and 10 ug sheared gDNA were used for each library preparation. Two PacBio SMRTbell™ libraries were size selected for fragments larger than 25 and 28 kb with the BluePippin™ device according to the manufacturer's instructions.

Size selected libraries were loaded with 28 to 80 pM on plate. Sequel II polymerase 2.0 was used in combination with the v4 PacBio sequencing primer and the Sequel II sequencing kit 2.0, run time was 15 and 20 hours to increase the amount of continuous long reads. A total of 21x Sequel II SMRT cells (8M) was sequenced leading to 2,85 Tbp of unique insert reads which represent about 42x effective genome coverage.

Due to the predicted significantly larger genome size of *L. paradoxa* (~100 Gb), we decided to generate highly accurate long PacBio HiFi or circular consensus sequences (CCS) to decrease the required computational effort, capacity and time. Long insert libraries were prepared as recommended by Pacific Biosciences according to the 'Guidelines for preparing HiFi SMRTbell libraries using the SMRTbell Express Template Prep Kit 2.0 (PN 101-853-100). In summary, long gDNA was sheared to 20 kb fragments with the MegaRuptor™ device (Diagenode) and 10 ug sheared gDNA was used for library preparation. We prepared nine

PacBio SMRTbell™ libraries that were size-selected for fragments larger than 6 to 9 kb with a BluePippin™ device according to the manufacturer's instructions.

The size selected libraries were run with 40 to 105 pM on plate with the SEQUEL II sequencing kit 2.0 for 30 hours, Sequel II polymerase 2.0 was used in combination with the v4 PacBio sequencing primer. A total of 103 Sequel II SMRT cells run on the SEQUEL IIs at the Dresden Concept Genome Center and the NGS Competence Center Tübingen. Circular consensus sequences were called making use of the default SMRTLink tools and DeepConsensus software tool (doi: https://doi.org/10.1101/2021.08.31.458403). We generated a total of 2,199 Gbp of consensus sequences representing ~25x coverage of the South American Lungfish genome.

3D genome confirmation capturing HiC. Chromatin conformation capturing was done for both species making use of the ARIMA HiC+ Kit (Material Nr. A410110) and followed the user guide for animal tissues (ARIMA-HiC 2.0 kit Document Nr: A160162 v00). In brief, circa 50 mg flash-frozen powdered liver tissue was crosslinked chemically for each pull-down. The crosslinked genomic DNA was digested with the restriction enzyme cocktail consisting of four restriction enzymes, respectively. The 5'-overhangs are filled in and labeled with biotin. Spatially proximal digested DNA ends are ligated and finally the ligated biotin containing fragments are enriched and went for Illumina library preparation, which followed the ARIMA user guide for Library preparation using the Kapa Hyper Prep kit (ARIMA Document Part Number A160139 v00). A total of four DNA pull-down reactions leading to five Illumina library preparations has been done for *L. paradoxa* (one library per 17.4 Gb) and two DNA pull-down reactions resulting in five Illumina library preparations for *P. annectens* (one library per 8,1 Gb genome size). The barcoded HiC libraries were run on S4 flow cell of a NovaSeq6000 with 2x 150 cycles to a coverage of 60x per genome.

10x Genomics linked reads sequencing of *P. annectens.* Linked Illumina reads for the South African lungfish have been generated using the 10x Genomics Chromium™ genome application following the Genome Reagent Kit Protocol v2 (Document CG00043, Rev B, 10x Genomics, Pleasonton, CA). In brief, 10 ng of ultra-long genomic DNA was partitioned across 1 Million Gel bead-in-emulsions (GEMS) using a Chromium™ devise. Individual gDNA molecules were amplified in these individual GEMS in an isothermal incubation using primers that contain a specific 16bp 10x barcode and the Illumima® R1 sequence. After breaking the emulsions, pooled amplified barcoded fragments were purified, enriched and libraries were prepared for Illumina sequencing as described in the protocol. Pooled Illumina libraries were sequenced to 60x genome coverage on an Illumina NovaSeq S4 flow cell.

Assembly of two lungfish genomes

Protopterus annectens:

PacBio reads. The target coverage for long read sequencing was 60X. Supplementary table 10 summarizes statistics on all raw data that we collected, and all data used for the assembly, which is the raw data except all those reads that were < 5kb in length and only the longest reads from each well of a SMRT cell. The expected coverage is the sum total of all base pairs collected, divided by the *post hoc* genome size of the final primary assembly.

Supplementary table 10: Statistics of PacBio dataset

	Number of SMRT cells	Number of Reads (M)	Total Base Pairs (Gbp)	Estimated Coverage	Average Read Length (Kbp)	Longest Read (Kbp)	Finish Date
Raw	15	160.5	3519.2	86.8	21.9	397	May 2020
Filtered_1		115.1	2847.5	70.2	24.7		
Filtered_2		103.9	2808.5	69.3	27.0		

The Raw data set contains all PacBio subreads longer than 2000 bp. The Filtered_1 data set contains only statistics for the longest read of each Zero-mode waveguide (ZMW). The Filtered_2 data set, that was used for the assembly, contains only the longest subread per ZMW with a minimum length of 5 kb.

10x Genomics linked reads. The target coverage for 10X Genomics linked reads sequencing was 60X. With the linked reads we followed two aims: First, error polishing of the assembled contigs and second, scaffolding of the assembled contigs. Supplementary table 11 summarizes statistics on the 10X Illlumina sequencing data.

Supplementary table 11: Statistics of 10X dataset and yield over 100 kbp

	Number of Reads (M)	Total Base Pairs (Gbp)	Estimated Coverage	Molecule Length N50 (Kbp)	Finish Date
Raw	22,182.2	3349.5	82.6	169.3	October 2021
barcode removed		3072.2	75.8		
Molecule length >=100Kb	---	966.9	23.8	227.1	

Raw: raw Illumina sequencing statistics; barcode removed: Illumina read pair statistics after trimming off the 16 bp 10x and 7 bp Illumina barcodes from the R1 reads; Molecule length >100Kb: Statistics for linked-reads molecule lengths greater than 100Kb. 10X reads were aligned to the final assembly with longranger align[1] The tool bxcheck (https://github.com/pd3/bxcheck) was used to calculate the molecule length statistics.

HiC Illumina read pairs. The target coverage for HiC sequencing was 60X. Supplementary table 12 shows the HiC sequencing statistics.

Supplementary table 12: Statistics of HiC dataset

Number of cycles	Number of Reads (M)	Total Bases Pairs (Gbp)	Duplication Rate	Estimated Coverage	Finish Date
150 PE	20,786.7	3138.8	18.3	77.4	November 2021

The estimated coverage was computed by using the final assembly sizes (including gap size). The duplication rate was calculated by pairtools dedup (https://github.com/open2c/pairtools).

Genome Assembly. De novo genome assembly was performed with DAmar (https://github.com/MartinPippel/DAmar). This assembler is based on an improved MARVEL assembler (https://github.com/schloi/MARVEL, commit ID: 5e17326) and the integration of parts from the DAZZLER, DALIGNER, DAMASKER, DASCRUBBER, DAZZ_DB, (https://github.com/thegenemyers; and the DACCORD (version: 0.0.635) code base. To assemble the genome, we performed the following steps: setup, PacBio read patching, assembly and error-polishing.

a) Setup Phase

PacBio reads were filtered by choosing only the longest read of each zero-mode waveguide (ZMW) and requiring subsequently a minimum read length of 5 kb. The resulting 103 million reads (69X coverage) were stored in a DAZZLER database, which was split with a block size of 500Mb.

b) Read Patching

The patch phase detects and corrects read artefacts including missed adapters, polymerase strand jumps, chimeric reads, and long low-quality read segments that are the primary impediments to long contiguous assemblies. To this end, we first computed local alignments of all raw reads. Because local alignment computation is by far the most time- and storage-consuming part of the pipeline, we reduced runtime and storage by masking repeats in the reads as follows. First, low-complexity intervals, such as microsatellites or homopolymers, were masked with DBdust (https://github.com/thegenemyers/DAZZ_DB). Second, tandem repeats were masked by using datander and TANmask (https://github.com/thegenemyers/DAMASKER). Third, we used a read alignment step to detect highly abundant repeats. We applied the repeat masking strategy from https://dazzlerblog.wordpress.com/2016/04/01/detecting-and-soft-masking-repeats/. To this end, we first compared each Dazzler database block (500Mb) against itself (.012X vs .012X coverage) softmasking with the dust and tandem tracks, and created a repeat mask rep.1 with a coverage threshold of 10. Next, we compared every group of 10 consecutive blocks (.12X vs .12X coverage) against themselves softmasking with rep.1 and created a mask rep.10 with a

coverage threshold of 20. Afterwards we stopped the iterative masking approach and merged all repeat tracks. Supplementary table 13 lists an overview about the individual repeat tracks.

Supplementary Table 13: Overview of the different repeat tracks during the incremental masking approach in the read patching phase

Track name	Bases	masked Bases	Avg Length
dust	389,771,363,362	13.9%	3,884
tandem	465,127,888,977	16.6%	4,773
rep1	191,275,424,076	6.8%	3,645
rep10	65,295,949,859	2.3%	3,035
merged track	550,859,261,598	19.6%	4,666

The merged repeat track was used to compute the all-vs-all alignments of the raw reads.

The repeat masks were subsequently used to prevent k-mer seeding in repetitive regions when computing all local alignments between all reads. Then we applied LAfix to detect and correct read artefacts. In LAfix we disabled chimeric read detection in repeat regions to avoid an over-fragmentation of the reads. As the softmasking potentially prevents getting the correct alignments in highly abundant repeat regions.

Supplementary table 14 shows the loss of reads in the read patching step. The bulk of the loss of 32.4% of the bases are attributed to reads that were fully softmasked, preventing daligner to compute local alignments.

Supplementary table 14: **Read statistics of raw reads and patched reads**

DB	Reads	Bases	% Bases	Avg Length
all reads	103,931,674	2,808,508,614,669	100.0%	27,022
patched reads	69,438,566	1,899,432,900,137	67.6%	27,354

c) *De novo* assembly

In the assembly phase, we first calculated all overlaps between patched reads using the same alignment strategy of the patch phase. The subsequent steps of (i) computing a quality track for all reads, (ii) computing a detailed repeat mask, (iii) filtering overlap piles, (iv) computing the

overlap graph, (v) touring the overlap graph to obtain primary contigs, and (vi) base error correction to create consensus sequences for the primary contigs follow the steps of the original MARVEL assembly pipeline.

Error Polishing. The assemblies (primary and alternate) were further polished by using the raw PacBio reads and applying two rounds of gcpp (https://github.com/PacificBiosciences/gcpp) polishing. Gcpp decodes polished sequences in capitals, whereas unpolished sequences are represented in lower case bases. DAMAR contigs tend to end within large repeats, that could not always be fully polished. To facilitate the later scaffolding process, uncorrected contig ends that remained after the second polishing round were trimmed back.

To further increase the QV value and reduce the remaining length errors in homopolymer regions, 10x read clouds were used. 10x read clouds were mapped to the Gcpp-polished contigs by using the 10X Genomics Longranger align pipeline, that makes use of the barcode-aware mapping tool Lariat. For run time reasons the alignment pipeline was stopped after the _ALIGNER step and several preset run time requirements were increased (see longranger_align.json).

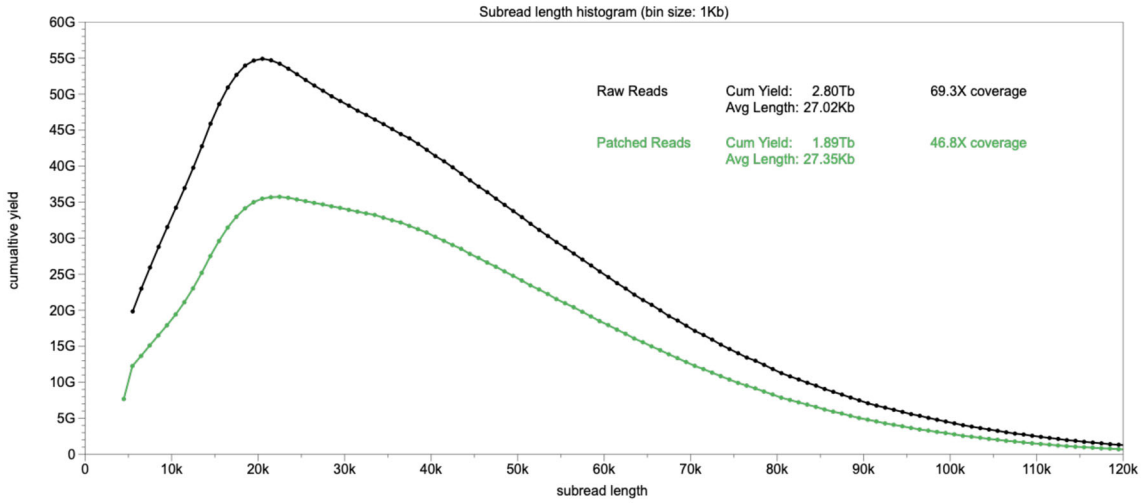
Afterwards the variant detector FreeBayes (version 1.3.4, default parameters + region argument to parallelise over number of contigs) detected polymorphic positions and fixed erroneous non-polymorphic sites in the reference sequence using bcftools consensus (version 1.14) (https://github.com/samtools/bcftools). 10x read cloud polishing was iteratively applied in two rounds.

Haplotype purging. Remaining haplotypic duplications in the primary contig set and alternate contig set were removed separately using purge dups (version 1.2.3[2]). To reduce the intermediate storage requirements in the self-alignment step we split the query-assembly into 20 chunks (3Gb each) and set the -I argument of minimap2 to 50G.

Scaffolding. The 10X Genomics read clouds were aligned to the primary contigs, and an adjacency matrix was computed from the barcodes using scaff10x. The alignment was done by using longranger align following the same strategy as described in the error polishing step. At the time of running the scaffolding, scaff10X (version 4.2) was not supporting more than 2 billion reads. The main developer of scaff10X, Zemin Ning, kindly supported us, created a code fix for the problematic step (scaff_barcode-sort) and provided us with a scaffold AGP file. The parameters that were used are: "-longread 1 -gap 100 -size 40.0 -matrix 2000 -reads 10 -score 20 -edge 50000 -link 8 -block 50000 -noalign align.dat tarseq.fastq fProAnn-scaff10x-hm-04.fa". Zemin Ning published his code changes in GitHub and scaff10X version 5.0 fully supports giant genomes and high read coverages.

For HiC-scaffolding we used yahs (version 1.1a) and the Dovetail Omni-C mapping pipeline (https://omni-c.readthedocs.io/en/latest/fastq_to_bam.html). Briefly, we mapped HiC reads using BWAMEM2 (version 2.2.1). Alignments were sorted, filtered and merged with the corresponding pairtools subcommands (version 0.3.0) and default parameter. The resulting deduplicated bam file together with the scaff10X AGP file were used as input for the yahs scaffolder.

Manual curation. We then performed a number of manual curation rounds to correct scaffolding errors, removed remaining haplotypes and to scaffold those contigs which were not automatically scaffolded into the 17 chromosomes. To this end, we developed a pipeline that automatically adapts the manual changes to the pairs file and scaffolds, (https://git.mpi-cbg.de/assembly/programs/manualcurationhic) and recreates the HiC density plots with pairtools and cooler. HiC maps were visually inspected with HiGlass (higlass-docker: version 0.9.0).



Supplementary Fig.1| Read length histogram of PacBio raw reads and PacBio patched reads. Due to the high repeat content, we lost 32.4% in the read patching step. The coverage dropped from 69X (raw reads) to 46.8X (patched reads).

Lepidosiren paradoxa

Consensus calling: Accurate consensus sequences for the 103 SMRT cells were created with PacBio's command line tool ccs (version 6.0.0, https://github.com/PacificBiosciences/ccs). We obtained 2199 Gb high quality CCS reads with a N50 of 12.99 Kb.

Genome Assembly: We ran HiFiasm (version 0.15.4-r347[3])to create the contig assembly with arguments: -l 2 --purge-cov 30 --primary. Remaining haplotypic duplications in the primary contig set were removed using purge_dups (version 1.2.3; [2]).

To polish the assembled contigs from the primary assembly, we mapped all CCS reads to the contig assemblies using pbmm2 (version 1.8.0) with arguments: --preset CCS -N 1 and called variants using DeepVariant (version 1.2.0). We then filtered for sites with 'genotype 1/1' to specify that all or nearly all reads support an alternative sequence at this position and a 'PASS' filter value to specify that the site passed DeepVariant's internal filters. We then corrected base errors using bcftools consensus consensus (version 1.14). Two rounds of polishing were applied. Merqury[4] (version 1.3) estimated a QV score of 68.1 when using PacBio CCS reads as underlying kmer database.

The assembly was checked for contaminations with blobtoolkit (version 1.1[5]) and an in-house pipeline which screens several blast databases.

Scaffolding. For scaffolding, we used yahs (version 1.1a) and the Dovetail Omni-C mapping pipeline (https://omni-c.readthedocs.io/en/latest/fastq_to_bam.html). Briefly, we mapped HiC reads using BWAMEM2 (version 2.2.1). Alignments were sorted, filtered, and merged with the corresponding pairtools subcommands (version 0.3.0) and default parameter. The following deduplication step had to be done on the uncompressed 12Tb pairsam file for performance reasons. The resulting deduplicated bam file was used as input for the yahs scaffolder.

We then performed a number of manual curation rounds to correct scaffolding errors, removed remaining haplotypes and to scaffold those contigs which were not automatically scaffolded into the 19 chromosomes. To this end, we developed a pipeline that automatically adapts the manual changes to the pairs file and scaffolds, (https://git.mpi-cbg.de/assembly/programs/manualcurationhic) and recreates the HiC density plots with pairtools and cooler. HiC maps were visually inspected with HiGlass (higlass-docker: version 0.9.0).

Genome size estimates. The genome size estimates through frequencies of *k*-mers revealed 91.2 Gb for the South American and 47.5 Gb for African lungfish with respective heterozygosities of 0.1% and 0.5%.

Transcriptome sequencing and assembly

Data collection and sequencing. For both, *Lepidosiren* and *Protopterus* total RNA was isolated from brain, gut, kidney, liver, lung, trunk muscle, pectoral fin and testis from the same individuals used to extract DNA for the genome assembly using a Qiagen RNeasy Mini Kit (Qiagen, Maryland, USA). RNA quality and quantity were assessed using a TapeStation 4150 system (Agilent Technologies, Palo Alto, CA) and a Qubit v4.0 fluorometer (Life Technologies, Darmstadt, Germany), respectively. For short-read sequencing, independent

libraries were constructed for each tissue (approximately 50 ng of total RNA as input) using a SENSE mRNA-Seq Library Prep Kit V2 (Lexogen, Vienna, Austria). Paired-end sequencing (2x150 bp) was then performed with an Illumina HiSeq X-Ten platform at the BGI Genomics Institute (Hong Kong). In total, 226.2 million (M) raw reads were obtained for *Lepidosiren*, ranging from 19.6 to 31.3M reads per tissue, while 193.8 M raw reads were obtained for *Protopterus*, ranging from 19.9 to 35.3 M reads per tissue. For long-read sequencing, RNA from each tissue was pooled equimolar in order to reach 10 µg and then a single library was constructed using a Direct cDNA Sequencing Kit (Oxford Nanopore, UK) following manufacturer's instructions. The final library constructed for each species was then sequenced in two flow cells (FLO-MIN106, R9.4.1) of a MinION Nanopore device. The two runs combined produced a total of 9.2 M raw reads with an average read length of 1,318 bp for *Lepidosiren*, and 7.9 M raw reads with an average read length of 1,210 bp for *Protopterus*.

Transcriptome assembly. Raw Illumina short-reads were filtered and corrected using Trimmomatic v0.39[6] and RCorrector v1.0.5[7] and processed for both *de novo* and reference-guided assembly. Raw Nanopore long-reads were processed with Pychopper v2 (https://github.com/epi2me-labs/pychopper) in default mode in order to identify, orient and trim full-length cDNA sequences. For *de novo* assembly, the Oyster River Protocol (ORP) v2.2.8[8] was used. Briefly, clean reads were assembled using Trinity v2.8.5[9] (k-mer=25), SPAdes v3.13.3[10] (k-mer=55), SPAdes (k-mer=75) and Trans-Abyss v2.0.1[11] (k-mer=32) respectively. The four different assemblies were then merged by the ORP's OrthoFuser module[12 13]. Completeness of the *de novo* assembled transcriptome was assessed with BUSCO v3[14] using the Core Vertebrate Genes (CVG) and the Vertebrata genes (vertebrata_odb9 database) in the gVolante webserver[15]. For reference-guided assembly, all clean short-reads of both lungfish species were aligned to the corresponding genome assembly, each sample independently, using HISAT2 v2.1.0[16] (maximum intron length set to 8 Mbp). The resulting mapping files, converted to BAM format and sorted by coordinates with Samtools v1.9[17], were parsed by StringTie v1.3.6[18] and transcripts reconstructed from each aligned sample were merged in a single consensus "gtf" file using the *merge* tool implemented in StringTie. Full-length Nanopore transcripts were aligned to the genomes using Minimap2 v2.23[19] and the resulting mapping file processed with StringTie. A final consensus reference-guided transcriptome was produced by combining the short- and long-read reconstructed transcripts using the StringTie *merge* module.

Genome Annotation

Genomes with large assembly and chromosomes challenge the genome annotation on multiple aspects: BLAST does not work with sequences larger than 1 Gb; coordinates larger than 2.2 Gb are not processable with int variables in C++ programs which are normally involved in genome annotation pipelines; AUGUSTUS hardly performs on large genes or long introns; and sequence mappers/aligners take a very long time to run. To overcome these problems, we devised a strategy based on shrinking the genome by retaining the entire coding region and small portions of their surrounding sequence space. To this end, first, we aligned the protein queries and transcripts to the unscaffolded contigs of the assembly. The mapped regions were then extended 600bp to both sides, merged and linked in order to scaffold contigs into chromosomes. After the annotation process, the gene coordinates on the shrunk assembly were lifted back to the original genome assembly

.

The genome was annotated used a previous pipeline[20,21] which was adapted for the very large genomes in this study.

The genome assembly was first screened to identify and mask repeat elements using RepeatModeler[22] and RepeatMasker (https://www.repeatmasker.org/). RepeatModeler identifies repeats ab initio and builds a *de novo* repeat library for the genome, which, together with Repbase[23] were used by RepeatMasker to identify and mask the repeats from the assembly. Given the large size of the genome, we ran this process on the assembly twice.

Protein coding genes were annotated by combining gene evidence from homology alignment, RNA/transcript mapping and *ab initio* prediction. For homology alignment, we first built a query library containing sequences of 406,436 proteins collected from UniProt/Swiss-prot, RefSeq (only "vertebrate_other") and NCBI genome annotations of *Homo sapiens*, *Danio rerio*, *Lepisosteus oculatus*, *Latimeria chalumnae*, *Xenopus tropicalis* and *Gallus gallus*. The queries were then aligned to the repeat soft-masked assembly using Exonerate (https://www.ebi.ac.uk/about/vertebrate-genomics/software/exonerate) and Genewise[24] For transcript alignment-based gene prediction. transcripts were reconstructed using the PASA pipeline v2.4.1[25] using as input either the *de novo* or the reference-guided transcriptome assembly (see previous section). *De novo* assembled transcripts were first processed by the seqClean tool (https://sourceforge.net/projects/seqclean) in order to remove poly-A tails and other contaminant sequences. PASA was run with the options "--transdecoder"; --max-intron-length 8000000" and using as transcript aligner the program Gmapl v2019-05-12[26]. The

obtained non-redundant alignment assemblies were used to generate an automated transcript-based genome annotation by applying the *pasa_asmbls_to_training_set.dbi* tool implemented in the PASA pipeline, with default parameters. The tool relies on TransDecoder v5.5.0 (https://github.com/TransDecoder) scripts in order to: 1) infer potential open reading frames (ORFs) in each assembled transcript (TransDecoder.LongOrfs) that are at least 100 amino acids long; 2) predict the likely coding regions (TransDecoder.Predict); 3) generate a genome-based coding region annotation file (cdna_alignment_orf_to_genome_orf.pl).

For *ab initio* gene prediction, AUGUSTUS[27] was first trained by good gene models and then ran with hints. Here, good gene models were those commonly predicted by all of the gene predictors based on homology alignment or RNA/transcript mapping, while hints refered to all of the homology and RNA based predictions.

The final consensus annotation was made by screening all predicted gene models throughout the assembly, selecting and modifying them by the following criteria: 1) when multiple homology gene models compete for a splice site, the one better supported by transcript evidence was chosen; 2) for a chosen gene model, its terminal exon (with start/stop codon) was replaced if a different terminal exon was predicted by AUGUSTUS and supported by transcript evidence; 3) an AUGUSTUS gene was kept when fully supported by transcript evidence and having no homology prediction competing for splice sites. The final gene set was screened by InterProscan (https://www.ebi.ac.uk/interpro/search/sequence/) to check for potential protein-domains, BUSCO[14] for completeness assessment, and Swiss-Prot & RefSeq BLAST[28,29] to assign gene symbols and names.

Genomic composition of repetitive elements

A total of 76.8% of the axolotl genome was annotated as repeated sequences. For the lungfish genomes those values are even higher, 80.8% in Australian lungfish, 84.2% in African lungfish, and 92.5% in South American lungfish (Extended Data Fig. 3a). While the genome size expansion in axolotl mainly resulted from accumulation of long terminal repeats (LTRs, 22.3%), in lungfish, the most abundant repeats are long interspersed elements (LINEs, 31.7%), suggesting that different expansion mechanisms occurred in the salamander and lungfish lineages. Yet, the content of unclassified repeats (“Unknown") is similar between salamanders and lungfish. The three lungfish lineages were enriched for different repeat TE superfamilies. In the African and Australian lungfish genomes LINE/CR1 (18.2%/16.3%) and LTR/DIRS (14.9%/10.7%) are the two repeat superfamilies forming the highest proportion, and

additionally LINE/L2 (9.2%) in Australian lungfish, while LINE/L1 is the most abundant repeat superfamily in the South American lungfish genome (14.2%), which suggest different genome expansion mechanisms even occurred among lungfish genomes (Extended Data Fig. 3b).

The historical TE expansion activity of all three lungfish and the axolotl lineages were analyzed using the Kimura 2-parameter distance of TE copies to their respective consensus. Interestingly, we found rather few TEs have accumulated on very recent timescales (close to 0% distance to consensus), suggesting that genome expansion happened over older evolutionary timescales. (Extended Data Fig. 5a).

Supplementary Information 2: Phylogenomics

Orthology inference

Predicted proteins from the new lungfish genomes genomes (*Protopterus annectens* and *Lepidosiren paradoxa*) were analyzed together with genome-predicted proteins from 17 representatives of major jawed vertebrate lineages, including lungfish (*Neoceratodus forsteri*), coelacanth (*Latimeria chalumnae*), ray-finned fishes (*Amphilophus citrinellus*, *Astatotilapia burtoni*, *Danio rerio*, *Lepisosteus oculatus*, *Tetraodon nigroviridis*, *Xiphophorus maculatus*), chondryichthyans (*Callorhinchus milii*), amphibians (*Ambystoma mexicanum*, *Xenopus laevis*), diapsids (*Anolis carolinensis*, *Gallus gallus*), and synapsids (*Homo sapiens*, *Mus musculus*). We selected phylogenetic hierarchical orthogroups at the tree root, as these have been shown to be more accurate[12]. From the initial 40,447 orthogroups, 12,497 contained data for all major sampled lineages (in practice, at least one sequence each for lungfishes or coelacanth, *Callorhinchus* or *Lepisosteus*, teleosts, amphibians, diapsids, and mammals). Homologous sets were aligned with MAFFT v7.304b[30] (default settings) and subjected to maximum likelihood inference with IQ-TREE 1.6.12[31] using fast searches, best-fit nuclear models selected using the Bayesian Information Criterion (BIC), and SH-like aLRT branch support (1000 pseudoreplicates) ('-m TEST -msub nuclear -fast -alrt 1000'). Multiple sequence alignments and trees for the 12,497 orthogroups we fed into phylopyrpruner v0.9.7 (https://pypi.org/project/phylopypruner/) to infer orthologs using a maximum inclusion criterion, removing too divergent sequences and OTUs with many paralogs on a per-locus basis ('--prune MI --mask pdist --trim-lb 5 --trim-freq-paralogs 4 --trim-divergent 1.25 --min-pdist 1e-8 --min-support 0.75 --min-taxa 6 --min-gene-occupancy 0.1 --min-otu-occupancy 0.1'). Phylopypruner inferred a total of 11,272 orthologs, of which 8,777 still had representatives for all major lineages (see taxonomic filter above). The 8,777 loci were then subjected to PREQUAL v1.02[32] to mask non-homologous amino acid stretches, aligned with MAFFT ginsi v7.304b with a variable scoring matrix ('--allowshift --unalignlevel 0.8') to avoid over-alignment of non-homologous regions[33], positions containing >75% gaps removed with ClipKIT v0.1 ('-m gappy')[34], and the resulting trimmed alignments concatenated into a single matrix.

Dataset assembly and quality-check

The phylogenetic utility of the 8,777 loci was assessed with genesortR[35] and the 5% most deviant loci were excluded, leaving 8,339. Visual inspection of the alignments and trees from loci showing the highest compositional heterogeneity (measured by RCFV[36] and highest

topological distance (lowest RF similarity) did not reveal obvious problems of contamination or paralogy. We then explored outlier loci by looking at the gene-wise log-likelihood difference in support for two competing relevant hypotheses (cf.[37]): lungfish + tetrapod vs. lungfish + coelacanth. To do so, we calculated per-site log-likelihood values for the two relevant topologies in IQ-TREE under BIC-selected model and used them to estimate per-gene log-likelihood differences, following[38]. Most loci displayed log-likelihood differences < 10 units (Fig. S1). We explored loci with the most extreme values (220 loci with log-likelihood differences > 10) by inspecting their alignments and trees for obvious cases of paralogy. Of these, 156 loci supported lungfish + tetrapod and 64 supported lungfish + coelacanth. We excluded 16 out of the latter 64 loci due to paralogy problems, leaving a final dataset of 8,323 loci that were concatenated into a single matrix.

Phylogeny inference. The phylogeny was inferred using PhyloBayes MPI v.1.9 [39] under the site-heterogeneous CAT model that can overcome phylogenetic artifacts such as long branch attraction when reconstructing basal sarcopterygian relationships[37]. Data were analyzed by gene jackknifing[40,41], i.e., creating 100 independent sets of loci each with at least 200,000 aligned amino acid positions. A total of 100 independent MCMC chains were run until convergence (20,000 cycles, saving every 10th cycle), assessed a posteriori using PhyloBayes' built-in functions (maxdiff = 1, meandiff = 0.00216271, ESS > 200 for all parameters after discarding the first 10% cycles as burnin). Post-burnin trees were summarized into a fully resolved consensus tree. Our results confirm the position of lungfishes as closest living relatives of tetrapods[37](Extended Data Fig. 1b,c) – in agreement with recent phylogenomic analyses based on RNAseq data[40]– now using a larger and more accurate set of orthologs inferred exclusively from genome data.

Molecular clock

Divergence times were inferred with a relaxed uncorrelated molecular clock, as implemented in MCMCTree within the PAML package v.4.9j[42]. We used a taxon-rich dataset consisting of 100 vertebrate taxa and 4,593 loci [40]. This allowed us to use 31 fossil calibrations following recent studies [40] [43,44] using uniform maximum and minimum ages as soft bounds.The uncorrelated clock model was selected following CorrTest, which did support rate autocorrelation in our dataset [45]. We used approximate likelihood calculations based on the gradient and Hessian matrix of the likelihood at the maximum likelihood estimates of branch lengths, calculated with CODEML (within the PAML package) under the best-fit JTT+Γ4

model. Priors on ancestral rate “rgene_gamma” were set to G(2, 11.19). Mean rates were approximated using the average root-to-tip paths in the PhyloBayes tree and a mean root age of 470 (mean between maximum-minimum bounds). The prior on the σ2 parameter (“sigma2_gamma”) was set to G(2,2) indicating substantial among-lineage rate heterogeneity. The tree prior assumed a uniform birth-death process with default parameters. The time unit was set to 100 Myr. Two independent MCMC chains were run for 200 million cycles, sampling every 10,000, after the initial 50,000 cycles that were discarded as burnin. Convergence was checked a posteriori in Tracer v.1.5 and all parameters obtained high ESS values >800.

Supplementary Information 3: Macrosynteny analysis

Chromosomes are remarkably well conserved among all three lungfish species investigated in this study. We have mapped genes that comprise the 24 bilaterian linkage groups (BLGs[46]) onto chromosomal scale assemblies of *Lepidosiren*, *Protopterus*, and *Neoceratodus* and confirmed the previously reported high degree of chromosomal synteny conservation[47] (Fig. 1, Extended Data Fig. 2). This observation is striking that in light of the substantial enlargement of the genomes, the chromosomal elements persisted. Only a few additional fusions could be detected in some lungfish genomes, e.g., chromosome 2 (scaf02) of *Lepidosiren* is a result of a recent fusion between chromosome 7 and parts of chromosome 1 of *Neoceratodus* (Fig. 1b). Most of the smaller chromosomes, however, remain fully conserved supporting the idea that microchromosomes evolve more slowly in the context of their syntenic representation [47] and represent ancestral building blocks of vertebrate karyotypes. In this context, we find recent fusions of microchromosomes into "newer" (and larger) chromosomes, e.g., *Protopterus* chromosome 23.24.25 (=NC_056747.1) is a result of the fusions of the more ancestral *Neoceratodus* chromosomes 23, 24, and 25 (Fig. 1). We also found examples of tetrapod-specific fusions, e.g., the homologs of Australian lungfish chromosomes 7, 13, 16, 23 fused with several other chromosomes in the tetrapod lineage, these four chromosomes thus represent the ancestral pre-lungfish/tetrapod condition. On the other hand, Australian lungfish chromosomes 11 and 17 (Fig. 1a) are the product of translocations that occurred before the separation of lungfishes and tetrapods (both chromosomes are retained within lungfishes and tetrapods but correspond to multiple chromosomes in the ray-finned fish outgroups).

Aside from this pattern, the chromosomes of the three species are largely collinear in their orthologous gene order (Extended Data Fig. 2). We have quantified collinearity by measuring uninterrupted "runs" of orthologous genes in the same order in a combination of two species. We find that the median length of perfectly collinear runs of orthologous genes comprises 13 genes between *Protopterus* and *Lepidosiren* and 8 genes between *Neoceratodus* and *Lepidosiren*. For comparison, in genomes where the gene order with homologous chromosomes has been fully scrambled, such as *Neoceratodus* and the chordate Amphioxus the median length of collinear run is 2 (full scrambling). This pattern has persisted for over ~150 million years of separate evolution since the extant lineages of lungfish last shared a common ancestor. While chromosome sizes are comparable across lungfishes, in cases of recent fusions (e.g., chromosome 8 in *Lepidosiren*, Fig. 1b), the overall collinearity pattern is not affected in the homologous regions of the chromosomes. Surprisingly, we found that microchromosomes more

often show breakages of collinearity than macrochromosomes [47]), in particular chromosome 15 of *Lepidosiren* and its homologous chromosome 5 in *Neoceratodus* are largely (completely) scrambled when compared to the otherwise highly collinear chromosomes 3 and 1, respectively (Extended Data Fig. 2). The observation that smaller chromosomes are conserved as whole syntenic units, yet the order of genes within them may be more scrambled, can be explained by elevated recombination rates on smaller chromosomes and a high repeat content.

Supplementary Information 4: Positive selection

Positively selected genes in all three extant lungfishes could be related to specific aspects of lungfish biology (Extended Data Fig. 7a). This includes a more terrestrially oriented lifestyle with dependence on air breathing, enhanced terrestrial olfaction, morphology of fins and their articulation. We also found support for proposed alterations in the hypothalamic-pituitary-thyroid axis related to neotenic aspects of lungfish biology. Furthermore, genes related to lungfish immunity (ETosis) and the challenges associated with managing a "giant genome" during essential processes such as cell division and transcription were found to be under positive selection. We also detect a large number of genes involved in the DNA damage response and apoptosis, which likely reflect the hyperactivity of transposons in the lungfishes` genomes.

Increased dependence on air breathing has resulted in cardiovascular and pulmonary adaptations in the lungfishes, including a partially divided heart, incipient double circulatory system and complex lungs [48]. In line with these adaptations towards a more terrestrial lifestyle, we detect positive selection on genes involved in cardiac and pulmonary ontogeny, homeostasis, and function as well as blood pressure regulation. These include key regulators of heart field and valve ontogeny (*foxc1a* [49], *nkx2.5* [50], *scx* [51], *smad6/7* [52,53]; genes associated with cardiac failure: (*s1pr1* [54,55]); lung, tracheal and alveolar ontogeny (*wnt7b* [56], *kcnj13* [57] and *etv5* [58]); genes involved in blood pressure regulation including adrenergic receptors (*adra2a, adra2db* and *adrb4c*) [59] and essential genes for lung function such as the cystic fibrosis associated gene *derl1* [60] and the formation of ciliated airways (*kif3* [61], *foxj1* [62]**).** The neuronal transcription factor *olig3*, which is specifically involved in the establishment of the hindbrain respiratory circuit and is related to hypoventilation in mammals [63]. Furthermore, the *shh* signal transduction cascade, essential for lung formation [64], is under positive selection with *shh* itself, its receptor *smo*, as well as other components (*hhati*, *uhl5* [65]). In relation to the general function of the cardiovascular system we detect positive selection on the blood coagulation cascade including *f7* [66] and *ptafr* [67].

Reliance on air breathing increased the scope and necessity for olfaction. We detect positive selection on genes related to olfactory neurogenesis (*eomes* [68], *gsx1* [69], *dmrta2* [70], *btbd3* [71]), olfactory biotransformation *(gstp1* [72]), olfactory cilia (*rp2* [73]) and odorant and vomeronasal receptors (*or2at4*, *or52k1*, *vmn2r1, vmn2r26*.)

A further class of genes under positive selection is related to fin/limb ontogenesis (*wnt10b* [74] *wnt2b* [75] and the genes of the *shh* pathway) and endoskeletal ossification (*itm2a* [76] , *dmrt2* [77]) as well as to the formation of joints, tendons and articular cartilages (*fmod, adamts4* [78], *has2* [79], *hs6st2* [80], *hyal2* [81], *crtap adamts4*, *mmp14* [82], *serpinh1* [83], *sost* [84]). Altogether this reflects the increase in number and importance of endochondral jointed fin elements in the lungfish lineage as well as the elaboration of synovial joints[85] towards the fin-to-limb transition. A further reason for positive selection on mineralization and endochondral ossification related genes could result from the secondary reduction in ossification in the lungfishes axial skeleton [86]. In relation to skeletal calcium homeostasis we detect selection on the parathyroid signalling (*casr, pth1r*, *vsp35l* [87]). In fish parathyroid hormones are secreted by the gills [88,89] and selection on the parathyroid pathway may reflect gill reductions as well as the secondary reduction in mineralised tissues [86]. In this context, the vestigial presence of gills in adult *Lepidosiren* may be solely related to their endocrine function since their respiratory capacity appears negligible [90].

Several authors have proposed that all three extant lungfishes are partially neotenic [86,91] explaining some of the anatomical features they share with salamanders, including their secondarily simplified brain architecture and the retention of external gills in *Protopterus*. In support of this hypothesis low levels of circulating T4 hormone and morphological alterations in their pituitary have been cited, indicating a disruption of the hypothalamic-pituitary-thyroid-axis (HPT axis) (reviewed [91]). Consistent with this hypothesis we detect positive selection on hormone receptors that regulate the endocrine function of the anterior pituitary through the hypothalamus (*gnrhr* [92], *prlhr* [93], *trhr* [94], *sstr1* [95]) and the melatonin hormone system (*asmt*, *mtnr1a*, *mtnr1bb*, *mtnr1* [96]). Furthermore we find positive selection on transcription factors involved in anterior pituitary ontogenesis (*pou2f1* [97], *hmx3* [98]). Altogether this provides further support for modifications in the lungfishes` HPT-axis potentially contributing to neotenic aspects of their morphology. Interestingly a correlation between genome size and paedomorphism has been described in urodeles[99], consistent with the occurrence of paedomorphism in lungfishes with giant genomes. We note that paedomorphism, which per definition involves the truncation of a developmental sequence, could evolve as the logical result of high mutagenic pressures as caused by overactivation of transposable elements. This could either occur via mutation of adult specific parts of the developmental program, or as an adaptive simplification of the life history stages to make the species more robust against such mutations.

For their innate immune response *Protopterus* have been shown to rely on the process of ETosis[100] (or NETosis), which involves the neutrophilic release of DNA to create extracellular traps for microorganisms[101] and genes involved in ETosis are under positive selection in lungfishes (*cxcr1*, *cxcr2*[102], *gzmb*[103], *prf* [103], *serpinb1*[104], *syk*[105]). We note that the importance of extracellular DNA traps in immunity provides adaptive significance to giant genomes containing enormous numbers of transposon DNA helping to ensnare microorganisms. Therefore, the evolution of a giant genome may be a form of transposon domestication related to innate immunity, instead of solely reflecting genomic hijacking by selfish elements due to compromised host immunity.

We find many cell cycle related genes under positive selection, particularly related to centrosome and spindle formation (*aurka-b*[106], *haus1*[107], *hepacam2*[108], *klhl13*[109], *limk2*, *sgo1*[110]), as well as to meiotic chromosome segregation (*ovol1*[111], *msh4*[112], dna replication (*meak7*[113], *plscr1*[114], *smarce1*[115] and dNTP precursor availability (*samhd1*[116]). We interpret this signal as being related to the challenges of having a "giant genome" which would be expected to require adaptive changes in the cell division machinery, to manage the replication and handling of the large amounts of DNA. Similar adaptations are likely required in the transcriptional machinery related to the presence of extremely long introns and we detect positive selection in genes related to initiation and processivity of all three RNA polymerases (*htatsf1*[117], *polr1f*[118], *nfia*[119]) as well as pre-mRNA splicing (*prpf18*[120]).

The hyperactivity of transposons is likely to induce a large degree of genotoxic stress in the germline. Indeed we find a large amount of genes to be under positive selection related to DNA damage response (e.g. *c1qbp, dcun1d3*[121], *eepd1*[122], *egln3*[123], *pum3*[124], *rnf144b*[125], *rnf19a*[126], *senp5*[127], *usp19*[128], *usp47*[129], *wdr25*[130], *wdr48*[131], *wwp1*[132], *xpa*[133], *znf341*[134]) genome integrity (*sirt7*[135], *smarca5*[135], *zbtb43*[136]) and apoptosis *(e.g,. aifm1*[137], *pdcd4*[138], *tmbim4*[139], *tnfaip3*[140] , *tradd*[141]). This signal is consistent with the huge amounts of DNA in the lungfishes' genomes and the expected genotoxic stress induced by elevated levels of transposon insertions. From the positive selection analysis alone, it is difficult to estimate whether these genes have evolved as part of an ongoing arms race between host and invading transposons to counteract transposition via somatic selection against damaged cells (i.e., apoptosis), or that alternatively this response has been partially muted leading to increased chances of survival of germline cells with DNA damage.

Supplementary Information 5: Gene losses

Genetic signature for attenuated DNA damage response and apoptosis

We analyzed gene losses across the lungfishes for genes associated with apoptosis and DNA damage. In the lineage of the *Lepidosirenidae* we detect gene losses related to these processes that suggest an attenuated response to genotoxic stresses (Extendend Data Fig. 7b). We interpret these losses as facilitating the expansion of the *Lepidosiren* and *Propoterus* genomes by reducing somatic selection on genotoxic stress induced by transposon insertion. *asb17* has been shown to specifically mediated sperm apoptosis upon genotoxic stress[142]. *rassf3* and *rassf6* are both mediators of apoptosis through *p53* following DNA damage[143-147]. *E2F8* is a known mediator of the DNA damage response including apoptosis[148]. *cab39l* and *stradb* are cofactors of *stk11* required for detection of dna damage and apoptosis[149,150], *rad51b* is involved in dna damage detection and repair by homologous recombination[151,152]. *dusp2*[153] and *mdp1*[154] are both phosphatases that dephosphorilate *stat3*, which in its phosphorylated state is an inhibitor of apoptosis by *p53*[155]. *htatip2* is a well-known tumor suppressor gene that facilitates apoptosis downstream of the DNA damage pathway[156]. Therefore, loss of these genes is predicted to act anti-apoptotically under conditions of genotoxic stress due to an attenuated DNA damage response. Such a cellular milieu reduces somatic selection against double stranded DNA breaks as induced by massive DNA transposition in the germline[157] (for instance resulting from reduced piRNA levels). Therefore, genome expansion – or alternatively the maintenance of extreme genome size – is facilitated in the *Lepidosirenidae* by reduced piRNA levels in combination with a reduced DNA damage response. It seems likely that transposon insertion itself was a causal mechanism for erosion of these genes. Under such a scenario it would have been the mutagenic pressure by selfish transposable elements to "hijack the cell" and produce a cellular environment that fuels their own propagation. Alternatively, these gene losses could be part of a process of transposon domestication favorable to the lungfishes. An adaptive significance of "giant genomes" composed mostly of transposons has not been established. It is intriguing that African Lungfishes have been shown to rely on neutrophilic cell lysis (ETosis), whereby extracellular DNA plays an essential role as microbial trap[100]. It is therefore likely that absolute secreted amounts of DNA quantitatively contribute to their immune defense, providing a strong selective incentive for a large genome size achieved through facilitating transposon activity.

Of note, *cab39l* has disappeared from all three lungfish species. Analysis of additional predicted gene losses in individual species or shared across the three lungfishes did not identify other

gene losses related to the DNA damage response and this signature is specific for the *Lepidosirenidae*.

Loss of Tetratricopetide Repeat Domain 23 (TTC23) indicating reduced sensitivity to hedgehog signalling

TTTC23 is a ciliary gene whose protein is present at the base of the cilia and which plays an important role in the intracellular transmission of the *hedgehog* signaling signal[158,159] **(**Extendend Data Fig. 7d**).** In vertebrate genomes two *TTC23* paralogs are present namely *TTC23* and *TTC23L* (*TTC23-like*) of which only *TTC23* has been functionally characterized and therefore the functional redundancy between these two proteins remains unknown, however loss of *TTC23* alone is sufficient to affect the sensitivity of cells to stimulation by the *hedgehog* agonist SAG[158] which activates the *hedgehog* receptor *smoothened*. In lungfishes, using BLAST we identified orthologs of *TTC23L* in *Lepidosiren*, *Protopterus* and *Neoceratodus*, but of only *TTC23* in *Neoceratodus*. Construction of a maximum likelihood tree including orthologs of human, mouse, coelacanth, spotted gar confirms the identity of the identified lungfish genes and suggests absence of *TTC23* in *Protopterus* and *Lepidosiren* (Extended Data Fig. 7d). We further investigated the inferred loss of *TTC23* in *Protopterus* and *Lepidosiren* by performing synteny analysis. *TTC23* is located in a highly conserved gene block in between *Synemin* (*SYNM*) and *Leucine Rich Repeat Containing 28* (*LRRC28*) and *Myocyte Enhancer Factor 2A* (*MEF2A)*. This gene block is present in *Protopterus* and *Lepidosiren* on chromosomes 7 and 15 respectively but without an annotated TTC23 gene. We analyzed the intervening region between SYNM and LRRC28 for human, mouse, spotted gar, *Neoceratodus*, *Protopterus* and *Lepidosiren* using pairwise Vista Lagan with translated anchoring[160] using the coelacanth sequence as baseline to verify the absence of conserved *TTC23* exons in this region in *Lepidosiren* and *Protopterus* (Extended Data Fig 7d). Whereas this analysis readily identifies the TTC23 exons in human, spotted gar and *Neoceratodus*, no homology is detected in *Lepidosiren* and *Protopteus*, further confirming the loss of *TTC23*.

The loss of *TTC23* has been shown to reduce cells responsiveness to *shh* signaling and provides a causative mechanism contributing to the reduction of both the paired fins and the scales in *Protopterus* and *Lepidosiren*. We note that loss of this gene would indicate a “ciliopathy” involved in the evolutionary modification of the appendages, analogous as to reported for the wing reduction that evolved in the Flightless cormorant (*Nannopterum harris*)[161].

Genetic signatures for scale and fin evolution

One of the genes reported to be lost by the ortho finder analysis is the signal transduction ligand *BMP3*, which acts as a negative regulator of BMP signaling by regulating receptor availability[162]. Synteny analysis of the *BMP3* locus in human, spotted gar and *Neoceratodus* indicates that *BMP3* is located in a highly conserved gene block in between *RASGEF1B*, *PRKG2* and *CFAP299*, *FGF5*, *PRDM8* and *ANTXR2* (Extended Data Fig. 7c). Search for the orthologs region in *Protopterus* and *Lepidosiren* identifies *ANTXR2* on chromosome 2 and 7 respectively but lacks the other genes in the same genomic context. Search for *RASGEF1B*, *PRKG2*, *FGF5* and *PRDM8* failed to identify their presence in *Lepidosiren* and *Protopterus*. The ciliary gene *CFAP299* was identified as a functional intron-less retrogene on chromosome 7 in *Protopterus* and on chromosome 15 in *Lepidosiren*. In addition, two pseudo retrogene copies of this gene are present in *Lepidosiren* on chromosomes 2 and 8. Altogether this indicates that the gene block including *RASGEF1B*, *PRKG2*, *BMP3*, *CFAP299*, *FGF5* and *PRDM8* was deleted in the common ancestor of *Protopterus* and *Lepidosiren* whereby the function of the original *CFAP299* gene is rescued by the presence of a retrogene copy. *PRKG2* deletion in mouse, humans and cattle[163-165] results in shorter long bones of the limbs and *RASGEF1B* is a downstream target of the shh pathway during limb development[166] and loss of these genes likely contributed to the evolution of a derived fin phenotype in *Lepidosirenidae* together with loss of *TTC23, hoxd12, e10, mm406* and regulatory evolution of the ZRS (see main text).

Mutation of *BMP3* has been shown to result in reduced squamation in zebrafish[167]. We further investigated expression of this gene during scale formation in the direct developing cichlid fish *Astatotilapia burtoni*[168] which confirms high expression during scale formation (Extended Data Fig. 7c). Therefore, loss of *BMP3* is likely to have contributed to the evolution of reduced squamation in the lineage of the *Lepidosirenidae*.

Supplementary Information 6: Genomic architecture and regulation of lungfish Hox clusters

Identification of lungfish hox clusters

Hox clusters were identified by BLASTN on the unassembled contigs (Supplementary table 15) in viroblast using the *Neoceratodus forsteri* hox genes as search query. For *Lepidosiren paradoxus* the *Hoxa*, *Hoxb* and *Hoxd* cluster are present on single contigs, while the Hoxc cluster is split on two contigs with a breakpoint between *hoxc11* and *hoxc12*. For *Protopterus annectens* the *Hoxa* and the *Hoxd* clusters were identified on single contigs, while the *Hoxb* cluster was present on two contigs with a breakpoint between *hoxb10* and *hoxb13*, and the *Hoxc* cluster was present on two contigs with a breakpoint between *hoxc11* and *hoxc12*. The previously published *Neoceratodus forsteri* genome[47] has contig level assembly of *Hoxa* and *Hoxc* clusters. Interestingly, all observed contig level breakpoints in any hox cluster in all available four lungfish genome assemblies occur in the AbdB (i.e. “posterior”) part of the cluster, either between *hoxa11* and *hoxa13*, *hoxb10* and *hoxb13*, *hoxc11* and *hoxc12* or *hoxd11/hoxd12* and *hoxd13*. This pattern likely reflects the loss of clustering between posterior and anterior hox genes and their drifting apart through the rise of an interspersed repeat desert. This splitting of an original cluster appears analogous to the situation in Drosophila where the ancestral Urbilaterian hox cluster has split into the anterior ANT-C and posterior BX-C complexes whereby distance constraints imposed by enhancer sharing preserve gene clustering only in selected parts of the cluster.

Supplementary Table 15: Contigs containing lungfish hox clusters. *Lepidosiren paradoxa* and *Protopterus annectens* contigs are from this study. *Neoceratodus forsteri* contigs are from[47].

Cluster/Species	Genes included	Contig
HoxA		
Lepidosiren paradoxa	*Hoxa1-Hoxa14*	ptg001146l_1
Protopterus annectens)	*Hoxa1-Hoxa14*	sc3_1243_02389_0_1_1_6059864
Neoceratodus forsteri	*Hoxa13-Hoxa14*	NFORS_037472_pilon_pilon
HoxB		
Lepidosiren paradoxa	*Hoxb1-Hoxb13*	ptg004017l_1
Protopterus annectens	*Hoxb1-Hoxb10*	sc13_1019_00466_0_1_1_555712
	Hoxb13	sc13_1020_14660_0_1_1_506349
Neoceratodus forsteri	*Hoxb1-Hoxb10*	NFORS_046705_pilon_pilon
	Hoxb13	NFORS_040414_pilon_pilon
HoxC		
Lepidosiren paradoxa	*Hoxc1-Hoxc10*	ptg004916l_1
	Hoxc11-Hoxc13	ptg002798l_1
Protopterus annectens	*Hoxc1-Hoxc10*	sc11_549_03762_0_1_1_5720896
	Hoxc11-Hoxc13	sc11_548_09420_0_1_1_1441400
Neoceratodus forsteri	*Hoxc1-Hoxc13*	NFORS_041640_pilon_pilon
HoxD		
Lepidosiren paradoxa	*Hoxd1-Hoxd13*	ptg007812l_1
Protopterus annectens	*Hoxd1-Hoxd13*	sc5_1178_01633_2_1_1_4613422
Neoceratodus forsteri	*Hoxd1-Hoxd12*	NFORS_044832_pilon_pilon
	Hoxd13	NFORS_035197_pilon_pilon

Hox genes are important developmental regulators that show evolutionary conserved clustering resulting from their complex joint regulation[169].We analyzed the topology of the four hox clusters in all the three lungfish species to determine how such spatially constrained loci evolve under conditions of genome expansion. The four lungfish hox clusters are overall expanded in all three species of lungfish; *Lepidosiren* possesses the largest clusters with respective sizes for the Hoxa cluster; 2.8Mb, Hoxb cluster; 3.7Mb, Hoxc cluster; 2.4Mb and Hoxd cluster; 2.0 Mb (Extended Data Fig. 8a, Extended Data Fig. 9). Compared to an approximate size of 0.1-0.2 Mb in mouse and coelacanth these clusters are approximately 20 times expanded and represent the largest known Hox clusters to date. In gene content the Lungfish clusters are similar to mouse, with the additional preservation of *hoxc1*, *hoxc3*, *hoxb10* and *hoxa14* (as in coelacanth) and the lineage specific loss of *hoxd12* in *Protopterus* and *Lepidosiren* (Extended Data Fig. 9**)** which is likely related to the reduction of their pectoral and pelvic fins. Expansion of the four hox clusters has not occurred homogenously but is confined to specific regions. Intronic

expansion has occurred in *hoxa3* and *hoxd3,* which have expanded approximately 100-fold to between 0.1 and 0.2 Mb (Extended Data Fig. 8a, Extended Data Fig. 9). Intergenic regions have expanded predominantly in the anterior and posterior parts of the cluster (Extended Data Fig. 8a, Extended Data Fig. 9). Most dramatic expansion has occurred between hox10/11 and hox12/13 genes in all four clusters, leading to a repeat rich gene dessert within the posterior cluster (Extended Data Fig. 8a, Extended Data Fig. 9). Strikingly, all four clusters possess a conserved core which is only marginally expanded compared to mouse and is low in repeat content compared to expanded regions (Extended Data Fig. 8a, Extended Data Fig. 9). This sub-clustering probably reflects purifying selection on transposons insertions caused by topology dependent regulatory constraints such as enhancer sharing and suggests that these sub-clusters remain jointly regulated. Therefore, genome expansion through repeat insertion is constrained in regions of functional gene clustering such as the hox clusters.

The *Hoxa* and *Hoxd* clusters are regulated by long range enhancers located 3`and 5`[170-172]The interactions with these regulatory landscapes are facilitated by a highly organized 3D chromatin conformation whereby the Hoxa and Hoxd clusters are located on the intersection of two abutting TADs (topologically associating domains)[171-173]. In line with their generally expanded genome sizes the synteny regions known to contain *Hoxa* and *Hoxd* regulatory elements are enlarged, raising the question how genome scaling influences the evolution of regulatory landscapes and their 3D chromatin organization. Evaluation of conserved Hoxa and Hoxd fin/limb enhancer sequences[170,171,173] across the three lungfishes show expected conservation profiles across the flanking synteny regions with the exception that hoxa enhancers mm406 and e10[172,173] appear gained in the lungfish/tetrapod branch but secondarily lost in *Lepidosirenidae*, possibly related to the reduction of their fins. The HiC dataset generated for *Protopterus*, together with a new dataset for Midas cichlid (*Amphilophus citrinellus*) allowed us to investigate how 3D genome interactions around the *Hoxa* and *Hoxd* clusters have evolved in expanded and shrunk synteny regions. Compared to human the synteny region surrounding the *Protopterus Hoxa* and *Hoxd* cluster is 5-10 times expanded whereas their Midas orthologous regions are about 10 times reduced in size. In spite of these size differences the HiC derived regulatory landscapes show a similar partitioning between two abutting TADs at the position of the Hoxa and Hoxd clusters altogether indicating that overall regulatory landscapes remain intact under evolving genome size and that the primary constraints on changing complex regulatory landscapes are imposed by enhancer sharing of clustered genes.

Appendix:

Neoceratodus forsteri and *Lepidosiren paradoxa* ZRS chemically synthesized fragment sequences. Grey highlight denotes vector cloning site sequences including a *Not*I restriction site (italic).

>NeoZRS
CTTCAGGCTGAAGCTGATGGAACA*GCGGCCGC*TTCTTGACACCAACCTGTGGATATGACGTT
ACTACATAAAATGTAACATTAATTTATCAGCGACAGCAACATCCTGAGCAATTATCCAAATT
ATCCAGACATCGCAAAATATTCCGTACAAGTGCAGTCTGTAGGATTTAGGAGTTTAACTCCT
TCAACATCAAAAGGGAAGCCTGATGGTAAAAAAATAAACAGTAGAAAAATTTTGAGGTAACT
TCCTTGCTTAATTAATTAGATGGACCAGGTGGAAGCGAAGAAGTCGGTGCTGGTACTCCAAA
TGTCTATAAAGCGAAACAACGTGACAGCACAATAGAGGAGGAACAAAGAGTTTTTTAATATG
CTTCTATCCTGTGTCACAGTTTGAAACTGTCCTGGTTTATGTCCCTTTTGGCAAACTTACAT
AAAAGTGACCCTGTACTGTATTTTATGACCAGATGACTTTTTCTGTGGCTAATTTGTATCAG
GCCCCATATTAAAGAGATGCAGAAATCCGTAGGAAGTACAAGGCTGTTTGTGTCCGTTACTT
TCATTGCATTCTTTCATTATATCTGCTCGTTTTTTTTCCCACTGATCATCCATAAATTGTTG
GAAATGAGTGATTGAGGAAGTGCTGCTCAGTGTTAGTTGCACATGCCTGCTCTGGGTATGTT
TTTTTGTGGGTGAGAGGAAATCATGCAACTGCACAAAGAAAAAGGGAAACTCCTGCTGGGAA
CCTTTCAAGGAAATTTAGCAGGCATAAAGGGGCTTGGTCTTGGTGTTTCACAGAAAAGAGCG
CATTATCTCTCCACCACCTCAGATTTTTAAGGGTGCCAACACTTTT*GCGGCCGCC*AGGAACA
TCCAAACTGAGCAGCC

>LepZRS
CTTCAGGCTGAAGCTGATGGAACA*GCGGCCGC*TTTCTTGACACCAACCTGTGGATATGATGG
TGCTACATAAAATGCAACTTTAATTTATCTGCGACAGCAACATCCTGAGCAATTATCCAAAT
TATCCAGACATTGCAAAATAATTCAGTACAAGTGTAGTCTGTAGGATTTAGGAGTTGAACTC
ATTCAACATCAAAAGGAAAGCTTGATAGTTAAAAAATAAACAGCAGAAAAAATTTGAGGTAA
CTTCCGTGTTTAATTAATTAGATGAACCAGGTGGAAGTGAAGAAGCTGGGTTGGTGCTCCCA
TAGTCTATAAAGCTAAACAACGTGACAGCACAATAGAAGAGGAACAAAGATTTTTTTAATAT
GTTTCTGTCCTGGGTCACAGTTTGAAATTGTCCTGGTTTATGTCCCTTGTGGCAAACTTGCA
TAAAAGTGATCTTGTACTGTATTTTATGACCAGATGACTTTTGCTGTGGCTAATTTGTATCA
GGCCTCATATTAAAAAGATGCAGAAGCCAACTTCTGTCTTGTTACTTTCATTGCATACTTTC
TTTATATATGCTCATTTTTCCCCCACTGATCATCCATAAATTGTTGGAAATGAGTGTTTAAG
GAAGTGCTGCTTAGTGTTATTTGCACATGCATGCTCTGGGTATGTTTTTTTGGTGATGAAAG
GAAATCATACAGCTGCAGAAAGAAAATTGGAAACTCCTGCTGGGAAGTTTTCAGGGAAATTT
AGCAGGCAGAAGGGTCTTGATTGTTTAACAGAGGCAAGCATGTTGTCGGTTTGTTGGCTTAT
TTGTTAAAGGTGCTAAAACTGTA*GCGGCCGC*CAGGAACATCCAAACTGAGCAGCC

References

1 Marks, P. *et al.* Resolving the full spectrum of human genome variation using Linked-Reads. *Genome Res* **29**, 635-645, doi:10.1101/gr.234443.118 (2019).

2 Guan, D. *et al.* Identifying and removing haplotypic duplication in primary genome assemblies. *Bioinformatics* **36**, 2896-2898, doi:10.1093/bioinformatics/btaa025 (2020).

3 Cheng, H., Concepcion, G. T., Feng, X., Zhang, H. & Li, H. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat Methods* **18**, 170-175, doi:10.1038/s41592-020-01056-5 (2021).

4 Rhie, A., Walenz, B. P., Koren, S. & Phillippy, A. M. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biol* **21**, 245, doi:10.1186/s13059-020-02134-9 (2020).

5 Challis, R., Richards, E., Rajan, J., Cochrane, G. & Blaxter, M. BlobToolKit - Interactive Quality Assessment of Genome Assemblies. *G3 (Bethesda)* **10**, 1361-1374, doi:10.1534/g3.119.400908 (2020).

6 Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114-2120, doi:10.1093/bioinformatics/btu170 (2014).

7 Song, L. & Florea, L. Rcorrector: efficient and accurate error correction for Illumina RNA-seq reads. *Gigascience* **4**, 48, doi:10.1186/s13742-015-0089-y (2015).

8 MacManes, M. D. The Oyster River Protocol: a multi-assembler and kmer approach for de novo transcriptome assembly. *PeerJ* **6**, e5428, doi:10.7717/peerj.5428 (2018).

9 Grabherr, M. G. *et al.* Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol* **29**, 644-652, doi:10.1038/nbt.1883 (2011).

10 Chikhi, R. & Medvedev, P. Informed and automated k-mer size selection for genome assembly. *Bioinformatics* **30**, 31-37, doi:10.1093/bioinformatics/btt310 (2014).

11 Robertson, G. *et al.* De novo assembly and analysis of RNA-seq data. *Nat Methods* **7**, 909-912, doi:10.1038/nmeth.1517 (2010).

12 Emms, D. M. & Kelly, S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol* **20**, 238, doi:10.1186/s13059-019-1832-y (2019).

13 Li, W., Jaroszewski, L. & Godzik, A. Clustering of highly homologous sequences to reduce the size of large protein databases. *Bioinformatics* **17**, 282-283, doi:10.1093/bioinformatics/17.3.282 (2001).

14 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, doi:10.1093/bioinformatics/btv351 (2015).

15 Nishimura, O., Hara, Y. & Kuraku, S. gVolante for standardizing completeness assessment of genome and transcriptome assemblies. *Bioinformatics* **33**, 3635-3637, doi:10.1093/bioinformatics/btx445 (2017).

16 Kim, D., Langmead, B. & Salzberg, S. L. HISAT: a fast spliced aligner with low memory requirements. *Nat Methods* **12**, 357-360, doi:10.1038/nmeth.3317 (2015).

17 Danecek, P. *et al.* Twelve years of SAMtools and BCFtools. *Gigascience* **10**, doi:10.1093/gigascience/giab008 (2021).

18 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, doi:10.1038/nprot.2016.095 (2016).

19 Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094-3100, doi:10.1093/bioinformatics/bty191 (2018).

20 Du, K. *et al.* Genome biology of the darkedged splitfin, Girardinichthys multiradiatus, and the evolution of sex chromosomes and placentation. *Genome Res* **32**, 583-594, doi:10.1101/gr.275826.121 (2022).
21 Du, K. *et al.* The sterlet sturgeon genome sequence and the mechanisms of segmental rediploidization. *Nat Ecol Evol* **4**, 841-852, doi:10.1038/s41559-020-1166-x (2020).
22 Flynn, J. M. *et al.* RepeatModeler2 for automated genomic discovery of transposable element families. *Proc Natl Acad Sci U S A* **117**, 9451-9457, doi:10.1073/pnas.1921046117 (2020).
23 Bao, W., Kojima, K. K. & Kohany, O. Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mob DNA* **6**, 11, doi:10.1186/s13100-015-0041-9 (2015).
24 Birney, E., Clamp, M. & Durbin, R. GeneWise and Genomewise. *Genome Res* **14**, 988-995, doi:10.1101/gr.1865504 (2004).
25 Haas, B. J. *et al.* Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res* **31**, 5654-5666, doi:10.1093/nar/gkg770 (2003).
26 Wu, T. D. & Watanabe, C. K. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics* **21**, 1859-1875, doi:10.1093/bioinformatics/bti310 (2005).
27 Stanke, M. & Morgenstern, B. AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. *Nucleic Acids Res* **33**, W465-467, doi:10.1093/nar/gki458 (2005).
28 Camacho, C. *et al.* BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421, doi:10.1186/1471-2105-10-421 (2009).
29 Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nat Methods* **12**, 59-60, doi:10.1038/nmeth.3176 (2015).
30 Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol* **30**, 772-780, doi:10.1093/molbev/mst010 (2013).
31 Nguyen, L. T., Schmidt, H. A., von Haeseler, A. & Minh, B. Q. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol Biol Evol* **32**, 268-274, doi:10.1093/molbev/msu300 (2015).
32 Whelan, S., Irisarri, I. & Burki, F. PREQUAL: detecting non-homologous characters in sets of unaligned homologous sequences. *Bioinformatics* **34**, 3929-3930, doi:10.1093/bioinformatics/bty448 (2018).
33 Katoh, K. & Standley, D. M. A simple method to control over-alignment in the MAFFT multiple sequence alignment program. *Bioinformatics* **32**, 1933-1942, doi:10.1093/bioinformatics/btw108 (2016).
34 Steenwyk, J. L., Buida, T. J., 3rd, Li, Y., Shen, X. X. & Rokas, A. ClipKIT: A multiple sequence alignment trimming software for accurate phylogenomic inference. *PLoS Biol* **18**, e3001007, doi:10.1371/journal.pbio.3001007 (2020).
35 Mongiardino Koch, N. Phylogenomic Subsampling and the Search for Phylogenetically Reliable Loci. *Mol Biol Evol* **38**, 4025-4038, doi:10.1093/molbev/msab151 (2021).
36 Zhong, M. *et al.* Detecting the symplesiomorphy trap: a multigene phylogenetic analysis of terebelliform annelids. *BMC Evol Biol* **11**, 369, doi:10.1186/1471-2148-11-369 (2011).
37 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, doi:10.1093/sysbio/syw057 (2016).

38 Shen, X. X., Hittinger, C. T. & Rokas, A. Contentious relationships in phylogenomic studies can be driven by a handful of genes. *Nat Ecol Evol* **1**, 126, doi:10.1038/s41559-017-0126 (2017).
39 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, doi:10.1093/sysbio/syt022 (2013).
40 Irisarri, I. *et al.* Phylotranscriptomic consolidation of the jawed vertebrate timetree. *Nat Ecol Evol* **1**, 1370-1378, doi:10.1038/s41559-017-0240-5 (2017).
41 Delsuc, F., Brinkmann, H., Chourrout, D. & Philippe, H. Tunicates and not cephalochordates are the closest living relatives of vertebrates. *Nature* **439**, 965-968, doi:10.1038/nature04336 (2006).
42 Yang, Z. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol* **24**, 1586-1591, doi:10.1093/molbev/msm088 (2007).
43 Marjanović, D. The Making of Calibration Sausage Exemplified by Recalibrating the Transcriptomic Timetree of Jawed Vertebrates. *Front Genet* **12**, 521693, doi:10.3389/fgene.2021.521693 (2021).
44 Brownstein, C. D., Harrington, R. C. & Near, T. J. The biogeography of extant lungfishes traces the breakup of Gondwana. *Journal of Biogeography* **50**, 1191-1198, doi:https://doi.org/10.1111/jbi.14609 (2023).
45 Tao, Q., Tamura, K., F, U. B. & Kumar, S. A Machine Learning Method for Detecting Autocorrelation of Evolutionary Rates in Large Phylogenies. *Mol Biol Evol* **36**, 811-824, doi:10.1093/molbev/msz014 (2019).
46 Simakov, O. *et al.* Deeply conserved synteny and the evolution of metazoan chromosomes. *Sci Adv* **8**, eabi5884, doi:10.1126/sciadv.abi5884 (2022).
47 Meyer, A. *et al.* Giant lungfish genome elucidates the conquest of land by vertebrates. *Nature* **590**, 284-289, doi:10.1038/s41586-021-03198-8 (2021).
48 Fishman, A. P., DeLaney, R. G. & Laurent, P. Circulatory adaptation to bimodal respiration in the dipnoan lungfish. *J Appl Physiol (1985)* **59**, 285-294, doi:10.1152/jappl.1985.59.2.285 (1985).
49 Yue, Y. *et al.* The transcription factor Foxc1a in zebrafish directly regulates expression of nkx2.5, encoding a transcriptional regulator of cardiac progenitor cells. *J Biol Chem* **293**, 638-650, doi:10.1074/jbc.RA117.000414 (2018).
50 McElhinney, D. B., Geiger, E., Blinder, J., Benson, D. W. & Goldmuntz, E. NKX2.5 mutations in patients with congenital heart disease. *J Am Coll Cardiol* **42**, 1650-1655, doi:10.1016/j.jacc.2003.05.004 (2003).
51 Levay, A. K. *et al.* Scleraxis is required for cell lineage differentiation and extracellular matrix remodeling during murine heart valve formation in vivo. *Circ Res* **103**, 948-956, doi:10.1161/circresaha.108.177238 (2008).
52 Galvin, K. M. *et al.* A role for smad6 in development and homeostasis of the cardiovascular system. *Nat Genet* **24**, 171-174, doi:10.1038/72835 (2000).
53 Chen, Q. *et al.* Smad7 is required for the development and function of the heart. *J Biol Chem* **284**, 292-300, doi:10.1074/jbc.M807233200 (2009).
54 Cannavo, A. *et al.* β1-adrenergic receptor and sphingosine-1-phosphate receptor 1 (S1PR1) reciprocal downregulation influences cardiac hypertrophic response and progression to heart failure: protective role of S1PR1 cardiac gene therapy. *Circulation* **128**, 1612-1622, doi:10.1161/circulationaha.113.002659 (2013).
55 Boal, F. *et al.* Galanin Regulates Myocardial Mitochondrial ROS Homeostasis and Hypertrophic Remodeling Through GalR2. *Front Pharmacol* **13**, 869179, doi:10.3389/fphar.2022.869179 (2022).
56 Wang, Z., Shu, W., Lu, M. M. & Morrisey, E. E. Wnt7b activates canonical signaling in epithelial and vascular smooth muscle cells through interactions with Fzd1, Fzd10,

and LRP5. *Mol Cell Biol* **25**, 5022-5030, doi:10.1128/mcb.25.12.5022-5030.2005 (2005).
57 Yin, W. *et al.* The potassium channel KCNJ13 is essential for smooth muscle cytoskeletal organization during mouse tracheal tubulogenesis. *Nat Commun* **9**, 2815, doi:10.1038/s41467-018-05043-5 (2018).
58 Zhang, Z. *et al.* Transcription factor Etv5 is essential for the maintenance of alveolar type II cells. *Proc Natl Acad Sci U S A* **114**, 3903-3908, doi:10.1073/pnas.1621177114 (2017).
59 Talmud, P. J. *et al.* Variants of ADRA2A are associated with fasting glucose, blood pressure, body mass index and type 2 diabetes risk: meta-analysis of four prospective studies. *Diabetologia* **54**, 1710-1719, doi:10.1007/s00125-011-2108-6 (2011).
60 Sun, F. *et al.* Derlin-1 promotes the efficient degradation of the cystic fibrosis transmembrane conductance regulator (CFTR) and CFTR folding mutants. *J Biol Chem* **281**, 36856-36863, doi:10.1074/jbc.M607085200 (2006).
61 Giridhar, P. V. *et al.* Airway Epithelial KIF3A Regulates Th2 Responses to Aeroallergens. *J Immunol* **197**, 4228-4239, doi:10.4049/jimmunol.1600926 (2016).
62 You, Y. *et al.* Role of f-box factor foxj1 in differentiation of ciliated airway epithelial cells. *Am J Physiol Lung Cell Mol Physiol* **286**, L650-657, doi:10.1152/ajplung.00170.2003 (2004).
63 Liu, Z. *et al.* Control of precerebellar neuron development by Olig3 bHLH transcription factor. *J Neurosci* **28**, 10124-10133, doi:10.1523/jneurosci.3769-08.2008 (2008).
64 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, doi:10.1165/rcmb.2014-0132TR (2015).
65 Zhou, Z. *et al.* The deubiquitinase UCHL5/UCH37 positively regulates Hedgehog signaling by deubiquitinating Smoothened. *J Mol Cell Biol* **10**, 243-257, doi:10.1093/jmcb/mjx036 (2018).
66 Bernardi, F. & Mariani, G. Biochemical, molecular and clinical aspects of coagulation factor VII and its role in hemostasis and thrombosis. *Haematologica* **106**, 351-362, doi:10.3324/haematol.2020.248542 (2021).
67 Papakonstantinou, V. D., Lagopati, N., Tsilibary, E. C., Demopoulos, C. A. & Philippopoulos, A. I. A Review on Platelet Activating Factor Inhibitors: Could a New Class of Potent Metal-Based Anti-Inflammatory Drugs Induce Anticancer Properties? *Bioinorg Chem Appl* **2017**, 6947034, doi:10.1155/2017/6947034 (2017).
68 Sessa, A., Mao, C. A., Hadjantonakis, A. K., Klein, W. H. & Broccoli, V. Tbr2 directs conversion of radial glia into basal precursors and guides neuronal amplification by indirect neurogenesis in the developing neocortex. *Neuron* **60**, 56-69, doi:10.1016/j.neuron.2008.09.028 (2008).
69 Pei, Z. *et al.* Homeobox genes Gsx1 and Gsx2 differentially regulate telencephalic progenitor maturation. *Proc Natl Acad Sci U S A* **108**, 1675-1680, doi:10.1073/pnas.1008824108 (2011).
70 Yoshizawa, A. *et al.* Zebrafish Dmrta2 regulates neurogenesis in the telencephalon. *Genes Cells* **16**, 1097-1109, doi:10.1111/j.1365-2443.2011.01555.x (2011).
71 Matsui, A. *et al.* BTBD3 controls dendrite orientation toward active axons in mammalian neocortex. *Science* **342**, 1114-1118, doi:10.1126/science.1244505 (2013).
72 Schwartz, M. *et al.* Interactions Between Odorants and Glutathione Transferases in the Human Olfactory Cleft. *Chem Senses* **45**, 645-654, doi:10.1093/chemse/bjaa055 (2020).

73 Maurya, D. K., Berghard, A. & Bohm, S. A multivesicular body-like organelle mediates stimulus-regulated trafficking of olfactory ciliary transduction proteins. *Nat Commun* **13**, 6889, doi:10.1038/s41467-022-34604-y (2022).
74 Brunelle, P. *et al.* WNT10B variants in split hand/foot malformation: Report of three novel families and review of the literature. *Am J Med Genet A* **179**, 1351-1356, doi:10.1002/ajmg.a.61177 (2019).
75 Ng, J. K. *et al.* The limb identity gene Tbx5 promotes limb initiation by interacting with Wnt2b and Fgf10. *Development* **129**, 5161-5170, doi:10.1242/dev.129.22.5161 (2002).
76 Van den Plas, D. & Merregaert, J. In vitro studies on Itm2a reveal its involvement in early stages of the chondrogenic differentiation pathway. *Biol Cell* **96**, 463-470, doi:10.1016/j.biolcel.2004.04.007 (2004).
77 Ono, K. *et al.* Dmrt2 promotes transition of endochondral bone formation by linking Sox9 and Runx2. *Commun Biol* **4**, 326, doi:10.1038/s42003-021-01848-1 (2021).
78 Shu, C. C., Flannery, C. R., Little, C. B. & Melrose, J. Catabolism of Fibromodulin in Developmental Rudiment and Pathologic Articular Cartilage Demonstrates Novel Roles for MMP-13 and ADAMTS-4 in C-terminal Processing of SLRPs. *Int J Mol Sci* **20**, doi:10.3390/ijms20030579 (2019).
79 Matsumoto, K. *et al.* Conditional inactivation of Has2 reveals a crucial role for hyaluronan in skeletal growth, patterning, chondrocyte maturation and joint formation in the developing limb. *Development* **136**, 2825-2835, doi:10.1242/dev.038505 (2009).
80 Wang, W. *et al.* Down-regulated HS6ST2 in osteoarthritis and Kashin-Beck disease inhibits cell viability and influences expression of the genes relevant to aggrecan metabolism of human chondrocytes. *Rheumatology (Oxford)* **50**, 2176-2186, doi:10.1093/rheumatology/ker230 (2011).
81 Higuchi, Y. *et al.* Conditional knockdown of hyaluronidase 2 in articular cartilage stimulates osteoarthritic progression in a mice model. *Sci Rep* **7**, 7028, doi:10.1038/s41598-017-07376-5 (2017).
82 Taylor, S. H. *et al.* Matrix metalloproteinase 14 is required for fibrous tissue expansion. *Elife* **4**, e09345, doi:10.7554/eLife.09345 (2015).
83 Wilkinson, D. J. Serpins in cartilage and osteoarthritis: what do we know? *Biochem Soc Trans* **49**, 1013-1026, doi:10.1042/bst20201231 (2021).
84 Lewiecki, E. M. Role of sclerostin in bone and cartilage and its potential as a therapeutic target in bone diseases. *Ther Adv Musculoskelet Dis* **6**, 48-57, doi:10.1177/1759720x13510479 (2014).
85 Askary, A. *et al.* Ancient origin of lubricated joints in bony vertebrates. *Elife* **5**, doi:10.7554/eLife.16415 (2016).
86 Bemis, W. Paedomorphosis and the evolution of the Dipnoi. *Paleobiology* **10**, 293-307 (1984).
87 Sims, N. A. VPS35: Two Ways to Recycle the Parathyroid Hormone Receptor (PTH1R) in Osteoblasts. *EBioMedicine* **9**, 3-4, doi:10.1016/j.ebiom.2016.06.029 (2016).
88 Okabe, M. & Graham, A. The origin of the parathyroid gland. *Proc Natl Acad Sci U S A* **101**, 17716-17719, doi:10.1073/pnas.0406116101 (2004).
89 Wang, K. *et al.* African lungfish genome sheds light on the vertebrate water-to-land transition. *Cell* **184**, 1362-1376.e1318, doi:10.1016/j.cell.2021.01.047 (2021).
90 de Moraes, M. F. *et al.* Morphometric comparison of the respiratory organs in the South American lungfish Lepidosiren paradoxa (Dipnoi). *Physiol Biochem Zool* **78**, 546-559, doi:10.1086/430686 (2005).

91 Joss, J. M. Lungfish evolution and development. *Gen Comp Endocrinol* **148**, 285-289, doi:10.1016/j.ygcen.2005.10.010 (2006).
92 Sun, J. *et al.* Effect of hypothyroidism on the hypothalamic-pituitary-ovarian axis and reproductive function of pregnant rats. *BMC Endocr Disord* **18**, 30, doi:10.1186/s12902-018-0258-y (2018).
93 Hinuma, S. *et al.* A prolactin-releasing peptide in the brain. *Nature* **393**, 272-276, doi:10.1038/30515 (1998).
94 Matre, V. *et al.* Molecular cloning of a functional human thyrotropin-releasing hormone receptor. *Biochem Biophys Res Commun* **195**, 179-185, doi:10.1006/bbrc.1993.2027 (1993).
95 Lam, K. S. & Wong, R. L. Thyroid hormones regulate the expression of somatostatin receptor subtypes in the rat pituitary. *Neuroendocrinology* **69**, 460-464, doi:10.1159/000054450 (1999).
96 Ciani, E. *et al.* Effects of Melatonin on Anterior Pituitary Plasticity: A Comparison Between Mammals and Teleosts. *Front Endocrinol (Lausanne)* **11**, 605111, doi:10.3389/fendo.2020.605111 (2020).
97 de Moraes, D. C., Vaisman, M., Conceição, F. L. & Ortiga-Carvalho, T. M. Pituitary development: a complex, temporal regulated process dependent on specific transcriptional factors. *J Endocrinol* **215**, 239-245, doi:10.1530/joe-12-0229 (2012).
98 Wang, W., Grimmer, J. F., Van De Water, T. R. & Lufkin, T. Hmx2 and Hmx3 homeobox genes direct development of the murine inner ear and hypothalamus and can be functionally replaced by Drosophila Hmx. *Dev Cell* **7**, 439-453, doi:10.1016/j.devcel.2004.06.016 (2004).
99 Martin, C. C. & Gordon, R. Differentiation trees, a junk DNA molecular clock, and the evolution of neoteny in salamanders. *Journal of Evolutionary Biology* **8**, 339-354, doi:https://doi.org/10.1046/j.1420-9101.1995.8030339.x (1995).
100 Heimroth, R. D. *et al.* The lungfish cocoon is a living tissue with antimicrobial functions. *Sci Adv* **7**, eabj0829, doi:10.1126/sciadv.abj0829 (2021).
101 Grol, M. W. *et al.* Tendon and motor phenotypes in the Crtap(-/-) mouse model of recessive osteogenesis imperfecta. *Elife* **10**, doi:10.7554/eLife.63488 (2021).
102 Teijeira, Á. *et al.* CXCR1 and CXCR2 Chemokine Receptor Agonists Produced by Tumors Induce Neutrophil Extracellular Traps that Interfere with Immune Cytotoxicity. *Immunity* **52**, 856-871.e858, doi:10.1016/j.immuni.2020.03.001 (2020).
103 Agak, G. W. *et al.* Extracellular traps released by antimicrobial TH17 cells contribute to host defense. *J Clin Invest* **131**, doi:10.1172/jci141594 (2021).
104 Farley, K., Stolley, J. M., Zhao, P., Cooley, J. & Remold-O'Donnell, E. A serpinB1 regulatory mechanism is essential for restricting neutrophil extracellular trap generation. *J Immunol* **189**, 4574-4581, doi:10.4049/jimmunol.1201167 (2012).
105 Mukherjee, M., Lacy, P. & Ueki, S. Eosinophil Extracellular Traps and Inflammatory Pathologies-Untangling the Web! *Front Immunol* **9**, 2763, doi:10.3389/fimmu.2018.02763 (2018).
106 Willems, E. *et al.* The functional diversity of Aurora kinases: a comprehensive review. *Cell Div* **13**, 7, doi:10.1186/s13008-018-0040-6 (2018).
107 Lawo, S. *et al.* HAUS, the 8-subunit human Augmin complex, regulates centrosome and spindle integrity. *Curr Biol* **19**, 816-826, doi:10.1016/j.cub.2009.04.033 (2009).
108 Ozaki, Y. *et al.* Poly-ADP ribosylation of Miki by tankyrase-1 promotes centrosome maturation. *Mol Cell* **47**, 694-706, doi:10.1016/j.molcel.2012.06.033 (2012).
109 Sumara, I. *et al.* A Cul3-based E3 ligase removes Aurora B from mitotic chromosomes, regulating mitotic progression and completion of cytokinesis in human cells. *Dev Cell* **12**, 887-900, doi:10.1016/j.devcel.2007.03.019 (2007).

110 Zhang, Q. & Liu, H. Functioning mechanisms of Shugoshin-1 in centromeric cohesion during mitosis. *Essays Biochem* **64**, 289-297, doi:10.1042/ebc20190077 (2020).
111 Li, B. *et al.* Ovol1 regulates meiotic pachytene progression during spermatogenesis by repressing Id2 expression. *Development* **132**, 1463-1473, doi:10.1242/dev.01658 (2005).
112 Paquis-Flucklinger, V. *et al.* Cloning and expression analysis of a meiosis-specific MutS homolog: the human MSH4 gene. *Genomics* **44**, 188-194, doi:10.1006/geno.1997.4857 (1997).
113 He, Z., Houghton, P. J., Williams, T. M. & Shen, C. Regulation of DNA duplication by the mTOR signaling pathway. *Cell Cycle* **20**, 742-751, doi:10.1080/15384101.2021.1897271 (2021).
114 Wyles, J. P., Wu, Z., Mirski, S. E. & Cole, S. P. Nuclear interactions of topoisomerase II alpha and beta with phospholipid scramblase 1. *Nucleic Acids Res* **35**, 4076-4085, doi:10.1093/nar/gkm434 (2007).
115 Collins, N. *et al.* An ACF1-ISWI chromatin-remodeling complex is required for DNA replication through heterochromatin. *Nat Genet* **32**, 627-632, doi:10.1038/ng1046 (2002).
116 Franzolin, E. *et al.* The deoxynucleotide triphosphohydrolase SAMHD1 is a major regulator of DNA precursor pools in mammalian cells. *Proc Natl Acad Sci U S A* **110**, 14272-14277, doi:10.1073/pnas.1312033110 (2013).
117 Li, X. Y. & Green, M. R. The HIV-1 Tat cellular coactivator Tat-SF1 is a general transcription elongation factor. *Genes Dev* **12**, 2992-2996, doi:10.1101/gad.12.19.2992 (1998).
118 Yuan, X., Zhao, J., Zentgraf, H., Hoffmann-Rohrer, U. & Grummt, I. Multiple interactions between RNA polymerase I, TIF-IA and TAF(I) subunits regulate preinitiation complex assembly at the ribosomal gene promoter. *EMBO Rep* **3**, 1082-1087, doi:10.1093/embo-reports/kvf212 (2002).
119 Wang, Z., Bai, L., Hsieh, Y. J. & Roeder, R. G. Nuclear factor 1 (NF1) affects accurate termination and multiple-round transcription by human RNA polymerase III. *Embo j* **19**, 6823-6832, doi:10.1093/emboj/19.24.6823 (2000).
120 Horowitz, D. S. & Krainer, A. R. A human protein required for the second step of pre-mRNA splicing is functionally related to a yeast splicing factor. *Genes Dev* **11**, 139-151, doi:10.1101/gad.11.1.139 (1997).
121 Ma, T. *et al.* DCUN1D3, a novel UVC-responsive gene that is involved in cell cycle progression and cell growth. *Cancer Sci* **99**, 2128-2135, doi:10.1111/j.1349-7006.2008.00929.x (2008).
122 Nickoloff, J. A. *et al.* Metnase and EEPD1: DNA Repair Functions and Potential Targets in Cancer Therapy. *Front Oncol* **12**, 808757, doi:10.3389/fonc.2022.808757 (2022).
123 Xie, L. *et al.* PHD3-dependent hydroxylation of HCLK2 promotes the DNA damage response. *J Clin Invest* **122**, 2827-2836, doi:10.1172/jci62374 (2012).
124 Chang, H. Y. *et al.* hPuf-A/KIAA0020 modulates PARP-1 cleavage upon genotoxic stress. *Cancer Res* **71**, 1126-1134, doi:10.1158/0008-5472.Can-10-1831 (2011).
125 Ho, S. R., Mahanic, C. S., Lee, Y. J. & Lin, W. C. RNF144A, an E3 ubiquitin ligase for DNA-PKcs, promotes apoptosis during DNA damage. *Proc Natl Acad Sci U S A* **111**, E2646-2655, doi:10.1073/pnas.1323107111 (2014).
126 Zhu, Q. *et al.* RNF19A-mediated ubiquitination of BARD1 prevents BRCA1/BARD1-dependent homologous recombination. *Nat Commun* **12**, 6653, doi:10.1038/s41467-021-27048-3 (2021).

127 Jin, Z. L., Pei, H., Xu, Y. H., Yu, J. & Deng, T. The SUMO-specific protease SENP5 controls DNA damage response and promotes tumorigenesis in hepatocellular carcinoma. *Eur Rev Med Pharmacol Sci* **20**, 3566-3573 (2016).
128 Wu, M. *et al.* USP19 deubiquitinates HDAC1/2 to regulate DNA damage repair and control chromosomal stability. *Oncotarget* **8**, 2197-2208, doi:10.18632/oncotarget.11116 (2017).
129 Parsons, J. L. *et al.* USP47 is a deubiquitylating enzyme that regulates base excision repair by controlling steady-state levels of DNA polymerase β. *Mol Cell* **41**, 609-615, doi:10.1016/j.molcel.2011.02.016 (2011).
130 Peñalosa-Ruiz, G. *et al.* WDR5, BRCA1, and BARD1 Co-regulate the DNA Damage Response and Modulate the Mesenchymal-to-Epithelial Transition during Early Reprogramming. *Stem Cell Reports* **12**, 743-756, doi:10.1016/j.stemcr.2019.02.006 (2019).
131 Goncalves, J. M., Cordeiro, M. M. R. & Rivero, E. R. C. The Role of the Complex USP1/WDR48 in Differentiation and Proliferation Processes in Cancer Stem Cells. *Curr Stem Cell Res Ther* **12**, 416-422, doi:10.2174/1574888x12666170315104013 (2017).
132 Chen, J. *et al.* DNA damage induces expression of WWP1 to target ΔNp63α to degradation. *PLoS One* **12**, e0176142, doi:10.1371/journal.pone.0176142 (2017).
133 Kusakabe, M. *et al.* Mechanism and regulation of DNA damage recognition in nucleotide excision repair. *Genes Environ* **41**, 2, doi:10.1186/s41021-019-0119-6 (2019).
134 Cekic, S. *et al.* Increased radiosensitivity and impaired DNA repair in patients with STAT3-LOF and ZNF341 deficiency, potentially contributing to malignant transformations. *Clin Exp Immunol* **209**, 83-89, doi:10.1093/cei/uxac041 (2022).
135 Paredes, S. *et al.* The epigenetic regulator SIRT7 guards against mammalian cellular senescence induced by ribosomal DNA instability. *J Biol Chem* **293**, 11242-11250, doi:10.1074/jbc.AC118.003325 (2018).
136 Meng, Y. *et al.* Z-DNA is remodelled by ZBTB43 in prospermatogonia to safeguard the germline genome and epigenome. *Nat Cell Biol* **24**, 1141-1153, doi:10.1038/s41556-022-00941-9 (2022).
137 Bano, D. & Prehn, J. H. M. Apoptosis-Inducing Factor (AIF) in Physiology and Disease: The Tale of a Repented Natural Born Killer. *EBioMedicine* **30**, 29-37, doi:10.1016/j.ebiom.2018.03.016 (2018).
138 Matsuhashi, S., Manirujjaman, M., Hamajima, H. & Ozaki, I. Control Mechanisms of the Tumor Suppressor PDCD4: Expression and Functions. *Int J Mol Sci* **20**, doi:10.3390/ijms20092304 (2019).
139 Liu, Q. TMBIM-mediated Ca(2+) homeostasis and cell death. *Biochim Biophys Acta Mol Cell Res* **1864**, 850-857, doi:10.1016/j.bbamcr.2016.12.023 (2017).
140 Martens, A. & van Loo, G. A20 at the Crossroads of Cell Death, Inflammation, and Autoimmunity. *Cold Spring Harb Perspect Biol* **12**, doi:10.1101/cshperspect.a036418 (2020).
141 Bender, L. M., Morgan, M. J., Thomas, L. R., Liu, Z. G. & Thorburn, A. The adaptor protein TRADD activates distinct mechanisms of apoptosis from the nucleus and the cytoplasm. *Cell Death Differ* **12**, 473-481, doi:10.1038/sj.cdd.4401578 (2005).
142 Yang, G. *et al.* E3 Ubiquitin Ligase ASB17 Promotes Apoptosis by Ubiquitylating and Degrading BCLW and MCL1. *Biology (Basel)* **10**, doi:10.3390/biology10030234 (2021).
143 Donninger, H., Schmidt, M. L., Mezzanotte, J., Barnoud, T. & Clark, G. J. Ras signaling through RASSF proteins. *Semin Cell Dev Biol* **58**, 86-95, doi:10.1016/j.semcdb.2016.06.007 (2016).

144 Iwasa, H. *et al.* The RASSF6 tumor suppressor protein regulates apoptosis and the cell cycle via MDM2 protein and p53 protein. *J Biol Chem* **288**, 30320-30329, doi:10.1074/jbc.M113.507384 (2013).
145 Kudo, T. *et al.* The RASSF3 candidate tumor suppressor induces apoptosis and G1-S cell-cycle arrest via p53. *Cancer Res* **72**, 2901-2911, doi:10.1158/0008-5472.Can-12-0572 (2012).
146 Morishita, M. *et al.* Characterization of mouse embryonic fibroblasts derived from Rassf6 knockout mice shows the implication of Rassf6 in the regulation of NF-κB signaling. *Genes Cells* **26**, 999-1013, doi:10.1111/gtc.12901 (2021).
147 Richter, A. M., Pfeifer, G. P. & Dammann, R. H. The RASSF proteins in cancer; from epigenetic silencing to functional characterization. *Biochim Biophys Acta* **1796**, 114-128, doi:10.1016/j.bbcan.2009.03.004 (2009).
148 Zalmas, L. P. *et al.* DNA-damage response control of E2F7 and E2F8. *EMBO Rep* **9**, 252-259, doi:10.1038/sj.embor.7401158 (2008).
149 Baas, A. F. *et al.* Activation of the tumour suppressor kinase LKB1 by the STE20-like pseudokinase STRAD. *Embo j* **22**, 3062-3072, doi:10.1093/emboj/cdg292 (2003).
150 Alessi, D. R., Sakamoto, K. & Bayascas, J. R. LKB1-dependent signaling pathways. *Annu Rev Biochem* **75**, 137-163, doi:10.1146/annurev.biochem.75.103004.142702 (2006).
151 Garcin, E. B. *et al.* Differential Requirements for the RAD51 Paralogs in Genome Repair and Maintenance in Human Cells. *PLoS Genet* **15**, e1008355, doi:10.1371/journal.pgen.1008355 (2019).
152 Lee, P. S. *et al.* RAD51B Activity and Cell Cycle Regulation in Response to DNA Damage in Breast Cancer Cell Lines. *Breast Cancer (Auckl)* **8**, 135-144, doi:10.4137/bcbcr.S17766 (2014).
153 Lu, D. *et al.* The phosphatase DUSP2 controls the activity of the transcription activator STAT3 and regulates TH17 differentiation. *Nat Immunol* **16**, 1263-1273, doi:10.1038/ni.3278 (2015).
154 Zhu, J. *et al.* Magnesium-dependent Phosphatase (MDP) 1 is a Potential Suppressor of Gastric Cancer. *Curr Cancer Drug Targets* **19**, 817-827, doi:10.2174/1568009619666190620112546 (2019).
155 Al Zaid Siddiquee, K. & Turkson, J. STAT3 as a target for inducing apoptosis in solid and hematological tumors. *Cell Res* **18**, 254-267, doi:10.1038/cr.2008.18 (2008).
156 Yu, X., Li, Z. & Wu, W. K. TIP30: A Novel Tumor-Suppressor Gene. *Oncol Res* **22**, 339-348, doi:10.3727/096504015x14424348426116 (2014).
157 Farkash, E. A. & Luning Prak, E. T. DNA damage and L1 retrotransposition. *J Biomed Biotechnol* **2006**, 37285, doi:10.1155/jbb/2006/37285 (2006).
158 Breslow, D. K. *et al.* A CRISPR-based screen for Hedgehog signaling provides insights into ciliary function and ciliopathies. *Nat Genet* **50**, 460-471, doi:10.1038/s41588-018-0054-7 (2018).
159 Shamseldin, H. E. *et al.* The morbid genome of ciliopathies: an update. *Genet Med* **22**, 1051-1060, doi:10.1038/s41436-020-0761-1 (2020).
160 Brudno, M. *et al.* LAGAN and Multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA. *Genome Res* **13**, 721-731, doi:10.1101/gr.926603 (2003).
161 Burga, A. *et al.* A genetic signature of the evolution of loss of flight in the Galapagos cormorant. *Science* **356**, doi:10.1126/science.aal3345 (2017).
162 Kokabu, S. *et al.* BMP3 suppresses osteoblast differentiation of bone marrow stromal cells via interaction with Acvr2b. *Mol Endocrinol* **26**, 87-94, doi:10.1210/me.2011-1168 (2012).

163 Díaz-González, F. *et al.* Biallelic cGMP-dependent type II protein kinase gene (PRKG2) variants cause a novel acromesomelic dysplasia. *J Med Genet* **59**, 28-38, doi:10.1136/jmedgenet-2020-107177 (2022).
164 Kawasaki, Y. *et al.* Phosphorylation of GSK-3beta by cGMP-dependent protein kinase II promotes hypertrophic differentiation of murine chondrocytes. *J Clin Invest* **118**, 2506-2515, doi:10.1172/jci35243 (2008).
165 Koltes, J. E. *et al.* A nonsense mutation in cGMP-dependent type II protein kinase (PRKG2) causes dwarfism in American Angus cattle. *Proc Natl Acad Sci U S A* **106**, 19250-19255, doi:10.1073/pnas.0904513106 (2009).
166 Lewandowski, J. P. *et al.* Spatiotemporal regulation of GLI target genes in the mammalian limb bud. *Dev Biol* **406**, 92-103, doi:10.1016/j.ydbio.2015.07.022 (2015).
167 Li, C. *et al.* Genome sequences reveal global dispersal routes and suggest convergent genetic adaptations in seahorse evolution. *Nat Commun* **12**, 1094, doi:10.1038/s41467-021-21379-x (2021).
168 Woltering, J. M., Holzem, M., Schneider, R. F., Nanos, V. & Meyer, A. The skeletal ontogeny of Astatotilapia burtoni - a direct-developing model system for the evolution and development of the teleost body plan. *BMC Dev Biol* **18**, 8, doi:10.1186/s12861-018-0166-4 (2018).
169 Duboule, D. The rise and fall of Hox gene clusters. *Development* **134**, 2549-2560, doi:10.1242/dev.001065 (2007).
170 Montavon, T. *et al.* A regulatory archipelago controls Hox genes transcription in digits. *Cell* **147**, 1132-1145, doi:10.1016/j.cell.2011.10.023 (2011).
171 Andrey, G. *et al.* A switch between topological domains underlies HoxD genes collinearity in mouse limbs. *Science* **340**, 1234167, doi:10.1126/science.1234167 (2013).
172 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, doi:10.1371/journal.pbio.1001773 (2014).
173 Berlivet, S. *et al.* Clustering of tissue-specific sub-TADs accompanies the regulation of HoxA genes in developing limbs. *PLoS Genet* **9**, e1004018, doi:10.1371/journal.pgen.1004018 (2013).