Title: Caecilian genomes reveal molecular basis of adaptation and convergent evolution of limblessness in snakes and caecilians

Vladimir Ovchinnikov [1]*, Marcela Uliano-Silva [2]*, Mark Wilkinson [3], Jonathan Wood [2], Michelle Smith [4], Karen Oliver [4], Ying Sims [2], James Torrance [2], Alexander Suh [5,6], Shane A. McCarthy [2,7], Richard Durbin [2,7] and Mary J. O'Connell [1].

[1]Computational and Molecular Evolutionary Biology Group, School of Life Sciences, Faculty of Medicine and Health Science, University of Nottingham, NG7 2RD, UK

[2]Tree of Life Programme, Wellcome Sanger Institute, Wellcome Genome Campus, Cambridge CB10 1SA, UK

[3]Department of Life Sciences, The Natural History Museum, London SW7 5BD, UK

[4]Scientific Operations, Wellcome Sanger Institute, Wellcome Genome Campus, Cambridge CB10 1SA, UK

[5]School of Biological Sciences, University of East Anglia, NR4 7TU, Norwich, UK

[6]Department of Organismal Biology, Science for Life Laboratory, Uppsala University, SE-752 36, Uppsala, Sweden

[7]Department of Genetics, University of Cambridge, Cambridge CB2 3EH, UK

* both authors contributed equally.

Corresponding author/s:
Mary O'Connell <Mary.O'Connell@nottingham.ac.uk>
Richard Durbin <rd109@cam.ac.uk>

Keywords:
Caecilia, Amphibia, Vertebrate Comparative Genomics, Limblessness.

Abstract:

We present genome sequences for *Geotrypetes seraphini* (3.8Gb) and *Microcaecilia unicolor* (4.7Gb) caecilians, a limbless, mostly soil-dwelling amphibian clade with reduced eyes, and unique putatively chemosensory tentacles. We identify signatures of positive selection unique to caecilians in 1,150 orthogroups, with enrichment of functions for olfaction and detection of chemical signals. All our caecilian genomes are missing the ZRS enhancer of Sonic Hedgehog, shown by *in vivo* deletions to be required for limb development in mice and also absent in snakes, thus revealing a shared

molecular target implicated in the independent evolution of limblessness in snakes and caecilians.

Living amphibians, frogs, salamanders and caecilians, diverged since the Triassic. They, or their ancestors, survived all mass extinctions including the Permian-Triassic which obliterated most terrestrial vertebrates[1]. In our current extinction crisis, amphibians are amongst the most threatened groups. Undoubtedly reference quality genomes will aid in conservation, disease ecology and evolution, and breeding programs, yet they are amongst the most challenging of vertebrate genomes, due in part to their large and highly repetitive genomes[2,3]. Gymnophiona (caecilians) are the deepest diverging of the three extant amphibian orders and comprise ~215 species which are classified into 10 families. They display an often underappreciated diversity of life history traits; including oviparity with either aquatic larvae or direct development (with or without post hatching parental care and skin feeding[4]) and viviparity. The Rhinatrematidae, the deepest diverging (~225 MYA) of the ten caecilian families[5], is represented by the only previously published caecilian genome *Rhinatremata bivittatum*, which is 5.3 Gb in size and was sequenced by the Vertebrate Genomes Project (VGP)[6].

Reference genomes of *Geotrypetes seraphini* (Dermopdiidae) and *Microcaecilia unicolor* (Siphonopidae) were assembled according to the VGP's 6.7.P5.Q40.C90 metric standards, the same used for *Rhinatrema bivittatum* and other vertebrates[6]. The assemblies present respectively contig N50 20.6Mb and 3.6Mb, scaffold N50 272Mb and 376Mb, Phred-scaled base accuracy Q43 and Q37, and 99% and 97% of the assemblies were assigned to chromosome models (**Supplementary Table S1**). Manual curation was performed as in Howe et al.[7] resulting in 69 and 55 removals of misjoins, 122 and 84 joins automatically missed, and 18 and 0 removals of false duplications for *G. seraphini* and *M. unicolor* respectively (**Supplementary Figure S1**). Chromosomal units were identified and named by size. The final assembly sizes were 3.8Gb and 4.7Gb, respectively **(Supplementary Table S1**). Chromosome content and gene order are conserved to a remarkable extent across caecilian chromosomes, with large blocks of colinear synteny up to chromosome scale further conserved to anurans (common frog and toad) across more than 600 million years of evolution (**Figure 1)**.

Substantial proportions of the caecilian genomes consist of repeats: a total of 67.7%, 72.5% and 69.3% for *R. bivittatum*, *G. seraphini* and *M. unicolor* respectively (**Supplementary Table S2**). Class I transposable elements (TEs; retrotransposons) are around 20 times more abundant (in %bp) than Class II TEs (DNA transposons) and make up more than 30% of each caecilian genome. LINEs are the most abundant transposon type, followed by DIRSs. These relative proportions differ from those found in the large genomes of other amphibians: a genomic low-coverage shotgun analysis of the caecilian *Ichthyophis bannanicus* (genome size 12.2 Gb) revealed the prevalent presence of DIRSs followed by LINEs[8], while salamander genomes are dominated by LTRs, and DIRSs never surpass 7% of the genomes[2,9]. These results reinforce the notion that repeated instances of extreme TE accumulation in amphibians do not reflect a failure to control a specific type of TE[8].

Comparing the protein coding regions across 22 vertebrate genomes we identified a set of 31,385 orthogroups, of which 15,216 contained caecilian genes. We identified 265 gene families present across vertebrates but missing in amphibia, and an additional 260 orthogroups lost specifically in caecilians (**Supplementary Table S3**). In contrast, 1,150 orthogroups are present only in caecilians (**Supplementary Table S4**), and are enriched for functions such as olfaction and detection of chemical signals (p-value<0.01). At least 20% of these caecilian specific genes contained one of three protein domains (zf-C2H2, KRAB, 7tm_4). The 7tm_4 proteins are transmembrane olfactory receptors[10]; enrichment of this domain amongst the novel protein families in caecilians suggests an intense selective pressure on chemosensory perception at the origin of the caecilians, as they adapted to life underground with reduced vision and compensatory elaboration of chemosensory tentacles. Proteins containing zf-C2H2 and KRAB domains are known to have functions in regulating transcription, with zf-C2H2 containing proteins in humans shown to recognize more motifs than any of the other transcription factors combined. In addition, KRAB and zf-C2H2-containing proteins have been shown to bind currently active and ancient families of specific TEs (e.g. LINEs and LTRs/ERVs)[11]. The emergence of novel gene families with these functional capacities at the origin of caecilians may have contributed to the unique pattern of TE accumulation we observe in this group; further work is needed.

We performed a gene birth and death analysis using CAFE v5[12] on the remaining 13,541 orthogroups, examining the ancestral and extant caecilian nodes where possible. The majority of these (10,035) orthogroups were excluded from the birth and death analysis because they had no net change in gene family size between caecilian species and the ancestral amphibian node (8,065 orthogroups), or had insufficient sampling (1,970 orthogroups). We reconstructed ancestral states for the remaining 3,506 orthogroups (**Supplementary Table S5**). There were 156 orthogroups that were completely absent in *G. seraphini* and *M. unicolor* (most likely lost in their most recent common ancestor) (**Supplementary Table S3**). Only 13 orthogroups showed significant changes in caecilians (**Figure 2, Supplementary Table S6**), with 5 expansions at the ancestral caecilian node (ACN), and 3 at the internal caecilian node (ICN), of which one gene family is significantly expanded at both nodes. There are a total of three gene families with significant contractions, all of which are on the ACN. The gene families displaying significant expansions are: cytochrome P450 family 2 (ACN), these monooxygenases catalyse many reactions involved in metabolism of a large number of xenobiotics and endogenous compounds[13]; butyrophilin (BTN) family (ACN), involved in milk lipid secretion in lactation and regulation of the immune response[14]; tripartite motif (TRIM) family (ACN and ICN) involved in a broad range of biological processes that are associated with innate immunity[15]; and H2A and H2B histones (ICN), which together with H3 and H4 histones and DNA form a nucleosome[16]. In contrast, while immune function related butyrophilin and TRIM families have significant expansions at the ACN and/or ICN, both immunoglobulin heavy and light variable gene families have significant contractions at the ACN. The final family displaying significant contractions is gamma crystallin, a structural protein found largely in the nuclear region of the lens of the eye at very high concentrations[17]. Changes in these gene family repertoires may have contributed to the transition to a fossorial lifestyle and packaging of a large genome.

We assessed selective pressure variation on the lineages leading to each extant caecilian and the ancestral caecilians (ACN and ICN) as compared to all other vertebrates in our dataset. In total, we detected 453 orthologous families with evidence of positive selective pressure acting across these nodes (**Supplementary Table S7**). On the ACN there was no statistically significant GO enrichment across the positively selected genes. Examples of genes with signatures of positive selection are: FBN1 (under positive selection on both the ACN and the ICN), AGTPBP1, and CEP290 all of which are involved in eye morphogenesis[18–20]. On the ICN there was significant GO enrichment for intermediate filament cytoskeleton function (GO:0045111). A sample of the genes under positive selection follow (specific internal caecilian node/lineage implicated are shown in parenthesis): HESX1 (*M. unicolor* and *R. bivittatum*) required for the normal development of the forebrain, eyes and other anterior structures such as the olfactory placodes and pituitary gland[21]; NFE2L2 (*G. seraphini*), a transcription factor that plays a key role in the response to oxidative stress: binds to antioxidant response elements present in the promoter region of many cytoprotective genes, such as phase 2 detoxifying enzymes, promoting their expression, thereby neutralizing reactive electrophiles[22–25]; LGR4 (*R. bivittatum*) is involved in the development of the anterior segment of the eye[26] and is required for the development of various organs, including kidney, intestine, skin and reproductive tract[27,28]; COL9A3 (*M. unicolor*, and *R. bivittatum*) encodes a component of Collagen IX - a structural component of cartilage, intervertebral discs and the vitreous body of the eye[29,30]. In summary, the cohort of genes under positive selection does not yield statistically significant enrichment for biological processes and functions, but there are a number of genes implicated in organ (especially eye) development and morphogenesis.

Enhancer sequence conservation between vertebrates is favoured in developmental regulator genes. For example, the I12a enhancer element, located between homeobox bigenes Dlx1 and Dlx2, is known to be conserved from bony fish to mice[31]. Analysis of the ortholog of the I12a enhancer across the 22 vertebrate species confirms that it is easily identifiable and conserved in all vertebrates, including the three caecilians (**Figure 2**). Snakes contain a mutant form of an otherwise well-conserved enhancer element known as ZRS that when placed into mice produces a "serpentised" phenotype, directly implicating it in vertebrate limblessness[32]. The ZRS enhancer element is highly conserved and located within the LMBR1 intron between orthologous exons in vertebrates. However, the conserved ZRS element is absent in the three caecilian species. In contrast, ZRS is intact in limbless lizards where a more complex and lineage-specific route to limblessness has been proposed[33]. Here, the absence of ZRS in caecilians, and the functional work on the mutated form of ZRS in snakes, provides us with a common molecular target for the convergent loss of limbs in snakes and caecilians.

Methods

Sample collection

Genomes were produced from wild-caught animals that had been maintained in captivity for several years. Voucher specimens are at the Natural History Museum, London: *Geotrypetes seraphini* (MW 11051, from Kon, Cameroon), *Rhinatrema bivittatum* (MW11052) and *Microcaecilia unicolor* (MW11053), both from Camp Patawa, Kaw Mountains, French Guiana.

DNA preparation, Sequencing and optical mapping

All DNA extractions were from liver tissue using the Bionano Animal Tissue Plug preparation (https://bionanogenomics.com/wp-content/uploads/2018/02/30077-Bionano-Prep-Animal-Tissue-DNA-Isolation-Soft-Tissue-Protocol.pdf). Pacific Biosciences libraries were prepared with the Express Template Prep Kit 1.0 and Blue Pippin size selected. Pacific Biosciences CLR data was generated from 36 SMRTcells of *M. unicolor* and 6 SMRTcells of *G. seraphini* sequenced with the S/P2-C2/5.0 sequencing chemistry on the Pacific Biosciences Sequel machine. A further 5 SMRTcells of *G. seraphini* were sequenced with S/P3-C1/5.0-8M sequencing chemistry on a Pacific Biosciences Sequel II machine. The Hi-C libraries were created with a Dovetail Hi-C kit for *G. seraphini* and an Arima Genomics kit (version 1) for *M. unicolor* and sequenced on an Illumina HiSeq X. A 10X Genomic Chromium machine was used to create the linked-read libraries and sequenced on an Illumina HiSeq X. Optical maps were created for both species using a Bionano Saphyr instrument.

Genome assembly

Assembly for *Geotrypetes seraphini* and *Microcaecilia unicolor* was conducted mainly as for *Rhinatrema bivittatum* described in Rhie et al.[6] using four data types and the Vertebrate Genomes Project (VGP) assembly pipeline (version 1.6 for *G. seraphini* and version 1.5 for *M. unicolor*; **Supplementary Figure S2**). In brief, the Pacific Biosciences CLR data for each species was input to the diploid-aware long-read assembler FALCON and its haplotype-resolving tool FALCON-UNZIP[34]. The resulting primary and alternate assemblies of *M. unicolor* were input to Purge Haplotigs[35] and *G. seraphini* assemblies were input to Purge_dups[36] for identification and removal of remaining haplotigs. Next, both species' primary assemblies were subject to two rounds of scaffolding using 10X long molecule linked-reads and Scaff10X (https://github.com/wtsi-hpag/Scaff10X) and one round of Bionano Hybrid-scaffolding with pre-assembled Cmaps from 1-enzyme non-nicking (DLE-1) and the Solve Pipeline. The resulting scaffolds were then further scaffolded into chromosome-scale scaffolds using the Dovetail/Arima library Hi-C data for *G. seraphini/M. unicolor* and SALSA2[37]. The scaffolded primary assemblies plus the Falcon-phased haplotigs were then subjected to Arrow[38] polishing with the Pacbio reads and two rounds of short read polishing using the 10X Chromium linked reads, longranger align[39], freebayes[40] and consensus calling with bcftools[41] (further details can be found at Rhie et al.[6] and Suppl Fig 1). Assemblies were checked for contamination and were manually curated using gEVAL system[42], HiGlass[43] and PretextView (https://github.com/wtsi-hpag/PretextView) as described previously[7]. Mitochondria were assembled using mitoVGP[44]. Assemblies and full annotations are available on NCBI under the accession numbers GCF_902459505.1 and GCF_901765095.1. Raw reads statistics, accession numbers and software versions employed can be found at **Supplementary Table S8 A**, **B** and **C**.

Repeats prediction and annotation

All caecilians were submitted to homology-based and de novo approaches for repeat identification and annotation. A de novo library of repeats was created for each species using the RepeatModeler2 package[45]. This library was then combined with Repbase "Amphibia" library (release 26.04) forming the final library for each species. Each assembly was searched for repeats with RepeatMasker (http://www.repeatmasker.org/). Repeat landscape plots were created with perl scripts from the RepeatMasker package.

Genome annotation

The three caecilian genomes were annotated using the NCBI Eukaryotic Genome Annotation Pipeline which produces homology-based and *ab initio* gene predictions to annotate genes (including protein-coding and non-coding as lncRNAs, snRNAs), pseudo-genes, transcripts, and proteins (for details see Annotation HandBook https://www.ncbi.nlm.nih.gov/genbank/eukaryotic_genome_submission_annotation/). In brief, first repeats are masked with RepeatMasker (http://www.repeatmasker.org/) and Window Masker[46]. Next, transcripts, proteins and RNA-Seq from the NCBI database are aligned to the genomes using Splign[47] and ProSplign (https://www.ncbi.nlm.nih.gov/sutils/static/prosplign/prosplign.html). Those alignments are submitted to Gnomon (https://www.ncbi.nlm.nih.gov/genome/annotation_euk/gnomon/) for gene prediction. Models built on RefSeq transcript alignments are given preference over overlapping Gnomon models with the same splice pattern. **Supplementary Table S9** presents a summary of caecilian annotations and details can be found on NCBI at the accessions GCF_902459505.1, GCF_901765095.1.

Data Assembly for Comparative study

The Coding DNA sequences (CDSs) for 21 vertebrate species (**Supplementary Figure S3**) were downloaded from Ensembl release 100[48]. In those cases where a more contemporary version of the genome was available on RefSeq (Release 200)[49] we used the RefSeq genome and corresponding annotations (**Supplementary Table S9**). The longest canonical protein coding region for each gene was retained for further analysis.

Orthogroup prediction and gene birth and death analysis:

We identified 31,385 orthogroups for the 419,877 protein coding regions across 21 vertebrate species using OrthoFinder[50]. We extracted the corresponding uncontroversial species tree from timetree.org[51]. The phylogenetic distribution of the orthogroups revealed 1,150 were gained in caecilians, and 525 that were absent in all three caecilians. We used a phylostratigraphic approach to explore caecilian specific losses in the context of the vertebrate phylogeny. Information about species-specific losses elsewhere in the tree was not carried forward for further analysis. We parsed the orthogroups that lack caecilians in the following ways: (1) to identify orthogroups that lack representation across all amphibia: we identified orthogroups that contained at least two fish species and two tetrapod (non-amphibian) species - totalling 265 orthogroups, (2) to identify orthogroups that are absent only in caecilians: we extracted those orthogroups with least two fish species and two tetrapod species (including at least one frog species) - totalling 238 orthogroups, (3) to identify orthogroups that are present across amphibia and amniota but absent in caecilians: we extracted orthogroups containing two frog species and two amniota species - totalling 22 orthogroups. Orthogroups that did not satisfy these filters had patterns of loss that were spurious across vertebrata. Combining the set of orthogroups that contain caecilian representatives (13,541) plus those that passed our filters 1-3 above (525), produced our final set of 14,066 orthogroups for analysis in CAFE v5 with the lambda parameter estimated for each species[12]. Statistically significant contractions or expansions of gene families are detailed in the main text, and all expansions and contractions are provided in **Supplementary Table S5**.

Analysis of selective pressure variation

Our selective pressure variation analysis focussed on 3,236 single-copy and 9,690 multi-copy genes from our orthogroups. The ML method we employed requires a minimum of 7 species[52] thus we removed families that did not meet this criterion. The 9,690 multicopy genes could be broken down into the following cohorts based on the CAFE predictions: there were 5,993 orthogroups with species-specific duplications, after this filter 3,464 of which were designated SGOs and 2,529 as multi-copy gene orthogroups (MGOs). There were 6,226 (which includes the 2529 MGOs) that were divided into their constituent single-copy paralogous groups using UPhO[53]. Note species-specific gene duplications that were not specific to caecilians were removed. A total of 14,807 single-copy gene orthogroups were identified in this way. We used a range of different alignment methods (MAFFT[54], MUSCLE[55], and Prank[56]) on each gene family and used MetAl[57] to choose the best fitting alignment method per gene. The corresponding gene trees were reconstructed using IQtree[58]. Robinson-Foulds distances between gene trees and the species tree were estimated using Clann[59], and only those gene trees with zero distance were retained for further analysis, i.e. the gene and species tree were in full agreement thus minimising the risk of hidden paralogy in our single-copy gene orthogroups (SGOs). We assessed the selective pressure variation using codon based models of evolution in codeml[60] using Vespasian[61] across all resulting 2,047 SGOs that satisfied all of the range of criteria described above. All alignments for the selective pressure analyses are at DOI:10.5281/zenodo.5780326.

GO Enrichment Analysis

The GO terms were predicted for all caecilian CDSs using EGGnog with default parameters (eggnog-mapper.embl.de)[62], and GO term enrichment analysis was carried out using goatools[63].

Comparative analysis of homologous enhancer elements

The ZRS enhancer sequence was identified using the method in Kvon et al.[32]. The ZRS enhancer in mouse is located within an intron between exons 5 and 6 of the LMBR1 gene sequence (Gene ID: 105804842). In brief, the approach involved extracting the LMBR1 sequence from the genomes of each species in our sample set (**Supplementary Table S10**) and identifying the homologous intron sequence containing the ZRS sequence across all species. Using BLASTn[64] the ZRS region was readily identifiable across all 22 species. The level of sequence conservation was quantified between mouse ZRS and all other species **(Figure 2,** detailed alignment of the E1 element within ZRS **Supplementary Figure S4).** The ZRS sequence was also searched against the complete genomes of all three caecilians (to account for possible relocation of the enhancer) and we did not identify a ZRS-like sequence in an alternative location in the caecilian genomes. Using the same approach, we quantified the level of sequence conservation across our set of vertebrates for an additional enhancer, I12a (AF349438.2), located between the homeobox bigene cluster paralogs DLX1 and DLX2 (**Supplementary Table S11**). For *Crocodylus porosus* we used the region between METAP1D and DLX2 because the DLX1 gene was not annotated in this species.

Acknowledgements

MUS, YS, JW, JT, KO, MS are supported by Wellcome grant WT206194, SAM and RD are supported by Wellcome grant WT207492, MW thanks the Direction de l'Environment de l'Aménagement et du Logement and Le Comité Scientifique Régional du Patrimonie Naturel, French Guiana. MJO'C would like to thank the University of Nottingham for awarding funds to support this work. MJO'C and VO are grateful for access to the University of Nottingham's Augusta HPC service. For the purpose of open access, and as this research was funded in part by the Wellcome Trust [Grant number WT206194 and WT207492], the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.

Author contributions

MW supplied all biological samples and contributed to the interpretation of results. MS performed DNA extractions and optical mapping. KO coordinated the creation of sequencing libraries and genomic sequencing. SAM generated the genome assemblies. YS, JT and JW performed the manual curation of the assemblies. MUS performed repeat analyses and BUSCO synteny analyses**.** MJO'C and VO performed and interpreted the selective pressure analyses and birth and death analyses. MJO'C and VO carried out the comparative analysis of ZRS and I12a enhancer elements and interpreted results. RD supervised the genomics aspects of the project and MJO'C the comparative analyses. All authors contributed to writing the manuscript.

References:

1. Wake, D. B. & Vredenburg, V. T. Are we in the midst of the sixth mass extinction? A view from the world of amphibians. *Proc. Natl. Acad. Sci. U. S. A.* **105**, 11466–11473 (2008).
2. Nowoshilow, S. *et al.* The axolotl genome and the evolution of key tissue formation regulators. *Nature* **554**, 50–55 (2018).
3. Funk, W. C., Zamudio, K. R. & Crawford, A. J. Advancing Understanding of Amphibian Evolution, Ecology, Behavior, and Conservation with Massively Parallel Sequencing. in *Population Genomics: Wildlife* (eds. Hohenlohe, P. A. & Rajora, O. P.) 211–254 (Springer International Publishing, 2018).
4. Wilkinson, M., Sherratt, E., Starace, F. & Gower, D. J. A new species of skin-feeding caecilian and the first report of reproductive mode in Microcaecilia (amphibia: Gymnophiona: Siphonopidae). *PloS One* **8**, e57756 (2013).
5. Wilkinson, M., San Mauro, D., Sherratt, E. & Gower, D. J. A nine-family classification of caecilians (Amphibia: Gymnophiona). *Zootaxa* **2874**, 41–64 (2011).
6. Rhie, A. *et al.* Towards complete and error-free genome assemblies of all vertebrate

species. *Nature* **592**, 737–746 (2021).
7. Howe, K. *et al.* Significantly improving the quality of genome assemblies through curation. *GigaScience* **10**, giaa153 (2021).
8. Wang, J. *et al.* Gigantic Genomes Provide Empirical Tests of Transposable Element Dynamics Models. *Genomics Proteomics Bioinformatics* **19**, 123–139 (2021).
9. Sun, C. & Mueller, R. L. Hellbender genome sequences shed light on genomic expansion at the base of crown salamanders. *Genome Biol. Evol.* **6**, 1818–1829 (2014).
10. Buck, L. & Axel, R. A novel multigene family may encode odorant receptors: a molecular basis for odor recognition. *Cell* **65**, 175–187 (1991).
11. Najafabadi, H. S. *et al.* C2H2 zinc finger proteins greatly expand the human regulatory lexicon. *Nat. Biotechnol.* **33**, 555–562 (2015).
12. Mendes, F. K., Vanderpool, D., Fulton, B. & Hahn, M. W. CAFE 5 models variation in evolutionary rates among gene families. *Bioinformatics* **36**, 5516–5518 (2020).
13. Manikandan, P. & Nagini, S. Cytochrome P450 Structure, Function and Clinical Significance: A Review. *Curr. Drug Targets* **19**, 38–54 (2018).
14. Afrache, H., Gouret, P., Ainouche, S., Pontarotti, P. & Olive, D. The butyrophilin (BTN) gene family: from milk fat to the regulation of the immune response. *Immunogenetics* **64**, 781–794 (2012).
15. Ozato, K., Shin, D.-M., Chang, T.-H. & Morse, H. C. TRIM family proteins and their emerging roles in innate immunity. *Nat. Rev. Immunol.* **8**, 849–860 (2008).
16. Koyama, M. & Kurumizaka, H. Structural diversity of the nucleosome. *J. Biochem.* **163**, 85–95 (2018).
17. Vendra, V. P. R., Khan, I., Chandani, S., Muniyandi, A. & Balasubramanian, D. Gamma crystallins of the human eye lens. *Biochim. Biophys. Acta* **1860**, 333–343 (2016).
18. Stephenson, K. A. J. *et al.* A FBN1 variant manifesting as non-syndromic ectopia lentis with retinal detachment: clinical and genetic characteristics. *Eye* **34**, 690–694 (2020).
19. Chakrabarti, L. *et al.* The Purkinje cell degeneration 5J mutation is a single amino acid insertion that destabilizes Nna1 protein. *Mamm. Genome* **17**, 103–110 (2006).
20. Sheck, L. *et al.* Leber Congenital Amaurosis Associated with Mutations in CEP290, Clinical Phenotype, and Natural History in Preparation for Trials of Novel Therapies. *Ophthalmology* **125**, 894–903 (2018).
21. Dattani, M. T. *et al.* Mutations in the homeobox gene HESX1/Hesx1 associated with septo-optic dysplasia in human and mouse. *Nat. Genet.* **19**, 125–133 (1998).
22. Huang, H. C., Nguyen, T. & Pickett, C. B. Regulation of the antioxidant response element by protein kinase C-mediated phosphorylation of NF-E2-related factor 2. *Proc. Natl. Acad. Sci. U. S. A.* **97**, 12475–12480 (2000).
23. Eggler, A. L., Small, E., Hannink, M. & Mesecar, A. D. Cul3-mediated Nrf2 ubiquitination and antioxidant response element (ARE) activation are dependent on the partial molar volume at position 151 of Keap1. *Biochem. J.* **422**, 171–180 (2009).
24. Huppke, P. *et al.* Activating de novo mutations in NFE2L2 encoding NRF2 cause a multisystem disorder. *Nat. Commun.* **8**, 818 (2017).
25. Sanghvi, V. R. *et al.* The Oncogenic Action of NRF2 Depends on De-glycation by Fructosamine-3-Kinase. *Cell* **178**, 807-819.e21 (2019).
26. Siwko, S., Lai, L., Weng, J. & Liu, M. Lgr4 in ocular development and glaucoma. *J. Ophthalmol.* **2013**, 987494 (2013).
27. Hoshii, T. *et al.* LGR4 regulates the postnatal development and integrity of male reproductive tracts in mice. *Biol. Reprod.* **76**, 303–313 (2007).
28. Kinzel, B. *et al.* Functional roles of Lgr4 and Lgr5 in embryonic gut, kidney and skin development in mice. *Dev. Biol.* **390**, 181–190 (2014).
29. Olsen, B. R. Collagen IX. *Int. J. Biochem. Cell Biol.* **29**, 555–558 (1997).
30. He, Y. & Karsdal, M. A. Chapter 9 - Type IX Collagen. in *Biochemistry Of Collagens,*

Laminins And Elastin: Structure, Function And Biomarkers (eds. Karsdal, M. A., Leeming, D. J., Henriksen, K. & Bay-Jensen, A.-C.) 67–71 (Academic Press, 2016).

31. Plessy, C., Dickmeis, T., Chalmel, F. & Strähle, U. Enhancer sequence conservation between vertebrates is favoured in developmental regulator genes. *Trends Genet.* **21**, 207–210 (2005).
32. Kvon, E. Z. *et al.* Progressive Loss of Function in a Limb Enhancer during Snake Evolution. *Cell* **167**, 633-642.e11 (2016).
33. Roscito, J. G. *et al.* Convergent and lineage-specific genomic differences in limb regulatory elements in limbless reptile lineages. *Cell Rep.* **38**, 110280 (2022).
34. Chin, C.-S. *et al.* Phased diploid genome assembly with single-molecule real-time sequencing. *Nat. Methods* **13**, 1050–1054 (2016).
35. Roach, M. J., Schmidt, S. A. & Borneman, A. R. Purge Haplotigs: allelic contig reassignment for third-gen diploid genome assemblies. *BMC Bioinformatics* **19**, 460 (2018).
36. Guan, D. *et al.* Identifying and removing haplotypic duplication in primary genome assemblies. *Bioinformatics* **36**, 2896–2898 (2020).
37. Ghurye, J. *et al.* Integrating Hi-C links with assembly graphs for chromosome-scale assembly. *PLoS Comput. Biol.* **15**, e1007273 (2019).
38. Chin, C.-S. *et al.* Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nat. Methods* **10**, 563–569 (2013).
39. Bishara, A. *et al.* Read clouds uncover variation in complex regions of the human genome. *Genome Res.* **25**, 1570–1580 (2015).
40. Garrison, E. & Marth, G. Haplotype-based variant detection from short-read sequencing. Preprint at https://arxiv.org/abs/1207.3907 (2012).
41. Danecek, P. *et al.* Twelve years of SAMtools and BCFtools. *GigaScience* **10**, giab008 (2021).
42. Chow, W. *et al.* gEVAL - a web-based browser for evaluating genome assemblies. *Bioinformatics* **32**, 2508–2510 (2016).
43. Kerpedjiev, P. *et al.* HiGlass: web-based visual exploration and analysis of genome interaction maps. *Genome Biol.* **19**, 125 (2018).
44. Formenti, G. *et al.* Complete vertebrate mitogenomes reveal widespread repeats and gene duplications. *Genome Biol.* **22**, 120 (2021).
45. Flynn, J. M. *et al.* RepeatModeler2 for automated genomic discovery of transposable element families. *Proc. Natl. Acad. Sci. U. S. A.* **117**, 9451–9457 (2020).
46. Morgulis, A., Gertz, E. M., Schäffer, A. A. & Agarwala, R. WindowMasker: window-based masker for sequenced genomes. *Bioinformatics* **22**, 134–141 (2006).
47. Kapustin, Y., Souvorov, A., Tatusova, T. & Lipman, D. Splign: algorithms for computing spliced alignments with identification of paralogs. *Biol. Direct* **3**, 20 (2008).
48. Yates, A. D. *et al.* Ensembl 2020. *Nucleic Acids Res.* **48**, D682–D688 (2020).
49. O'Leary, N. A. *et al.* Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* **44**, D733-745 (2016).
50. Emms, D. M. & Kelly, S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol.* **20**, 238 (2019).
51. Kumar, S., Stecher, G., Suleski, M. & Hedges, S. B. TimeTree: A Resource for Timelines, Timetrees, and Divergence Times. *Mol. Biol. Evol.* **34**, 1812–1819 (2017).
52. Anisimova, M., Bielawski, J. P. & Yang, Z. Accuracy and power of bayes prediction of amino acid sites under positive selection. *Mol. Biol. Evol.* **19**, 950–958 (2002).
53. Ballesteros, J. A. & Hormiga, G. A New Orthology Assessment Method for Phylogenomic Data: Unrooted Phylogenetic Orthology. *Mol. Biol. Evol.* **33**, 2117–2134 (2016).
54. Rozewicki, J., Li, S., Amada, K. M., Standley, D. M. & Katoh, K. MAFFT-DASH: integrated protein sequence and structural alignment. *Nucleic Acids Res.* **47**, W5–W10 (2019).
55. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high

throughput. *Nucleic Acids Res.* **32**, 1792–1797 (2004).

56. Löytynoja, A. Phylogeny-aware alignment with PRANK. *Methods Mol. Biol. Clifton NJ* **1079**, 155–170 (2014).
57. Blackburne, B. P. & Whelan, S. Measuring the distance between multiple sequence alignments. *Bioinformatics* **28**, 495–502 (2012).
58. 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 (2015).
59. Creevey, C. J. & McInerney, J. O. Clann: investigating phylogenetic information through supertree analyses. *Bioinformatics* **21**, 390–392 (2005).
60. Xu, B. & Yang, Z. PAMLX: a graphical user interface for PAML. *Mol. Biol. Evol.* **30**, 2723–2724 (2013).
61. Constantinides, B. *et al.* Vespasian: genome scale detection of selective pressure variation (Version 0.5.3) [Computer software]. *GitHub* https://doi.org/10.5281/zenodo.5779868 (2021).
62. Huerta-Cepas, J. *et al.* eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res.* **47**, D309–D314 (2019).
63. Klopfenstein, D. V. *et al.* GOATOOLS: A Python library for Gene Ontology analyses. *Sci. Rep.* **8**, 10872 (2018).
64. Camacho, C. *et al.* BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421 (2009).

Figure 1: **Synteny plots showing the conservation of large scale gene linkage and gene order across caecilians, and to a substantial extent across amphibia.** Conserved unique single copy vertebrate genes were identified with BUSCO and connected by lines coloured according to their chromosomal location in *Rhinatrema bivittatum*. Common frog *Rana temporaria* and toad *Bufo bufo* genomes from https://wellcomeopenresearch.org/articles/6-286 and https://wellcomeopenresearch.org/articles/6-281 respectively. Synteny was created with ChrOrthLink (https://github.com/chulbioinfo/chrorthlink). Images of caecilians are modified using the Gimp software from original photos taken by Mark Wilkinson. Frog and Toad silhouettes are taken from http://phylopic.org/.

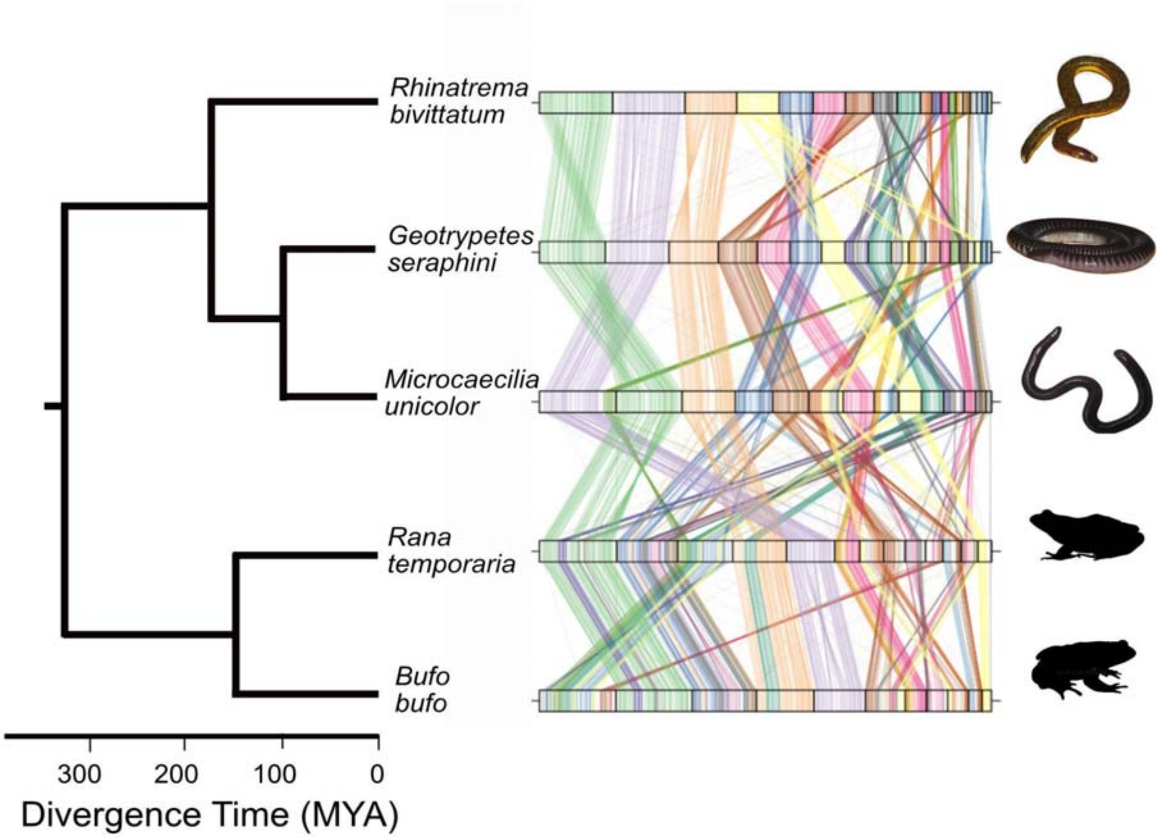
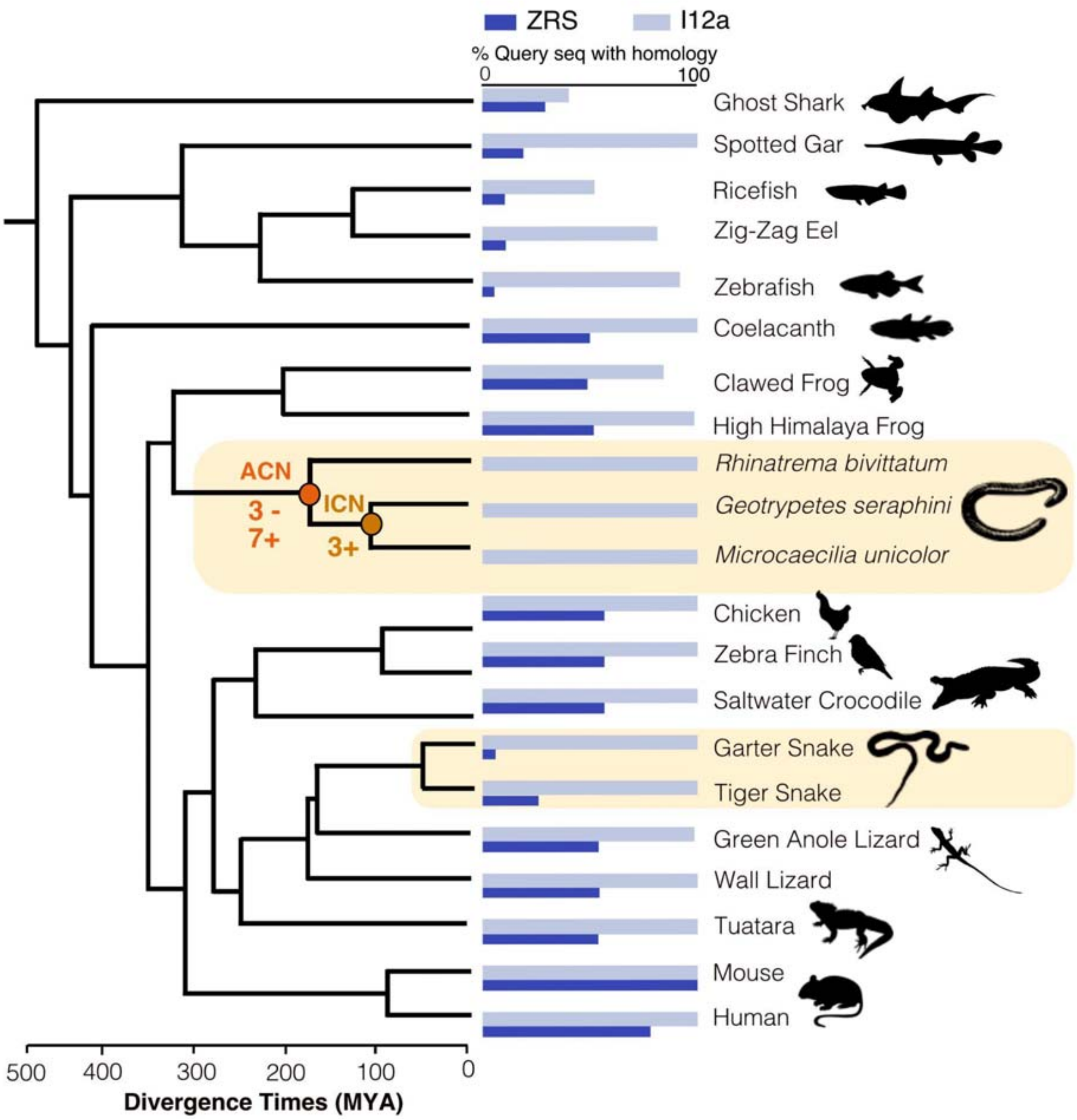


Figure 2: **Summary of gene gain and loss and levels of conservation of two enhancer elements across vertebrates (ZRS and I12a).** The vertebrate species phylogeny is shown on the left with the significant gene gain and loss events noted on the ancestral and internal caecilian nodes (ACN and ICN), respectively. The histogram shows the level of sequence conservation for each species for two enhancers: ZRS (dark blue) and I12a (light blue). Snakes and caecilians are highlighted as they independently evolved limbless morphologies. Animal images taken from http://phylopic.org/

.



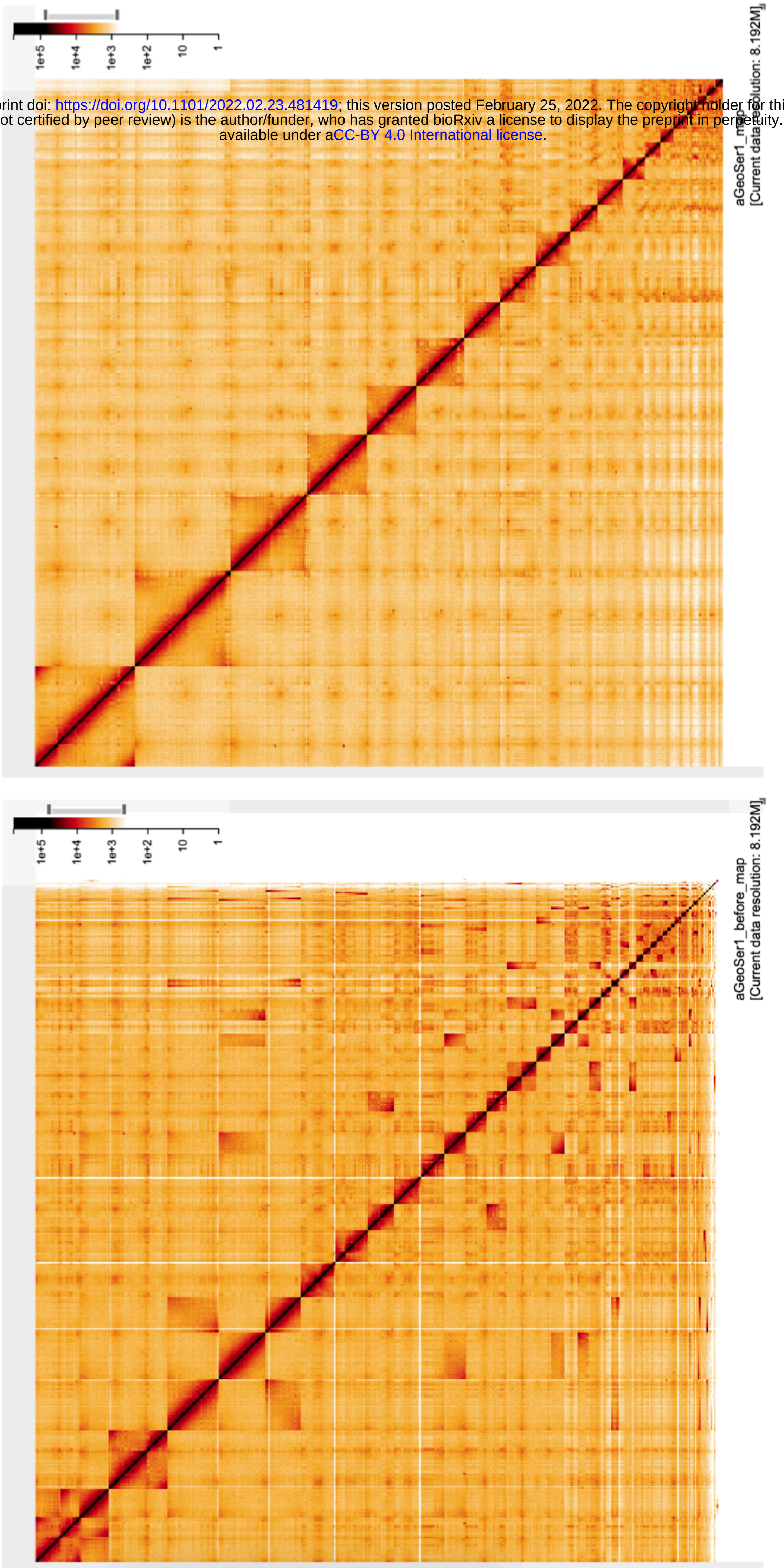
Supplementary Figures and legends:

Supplementary Figure S1: Hi-c heatmaps before (left) and after (right) manual curation for G. seraphini and M. unicolor.

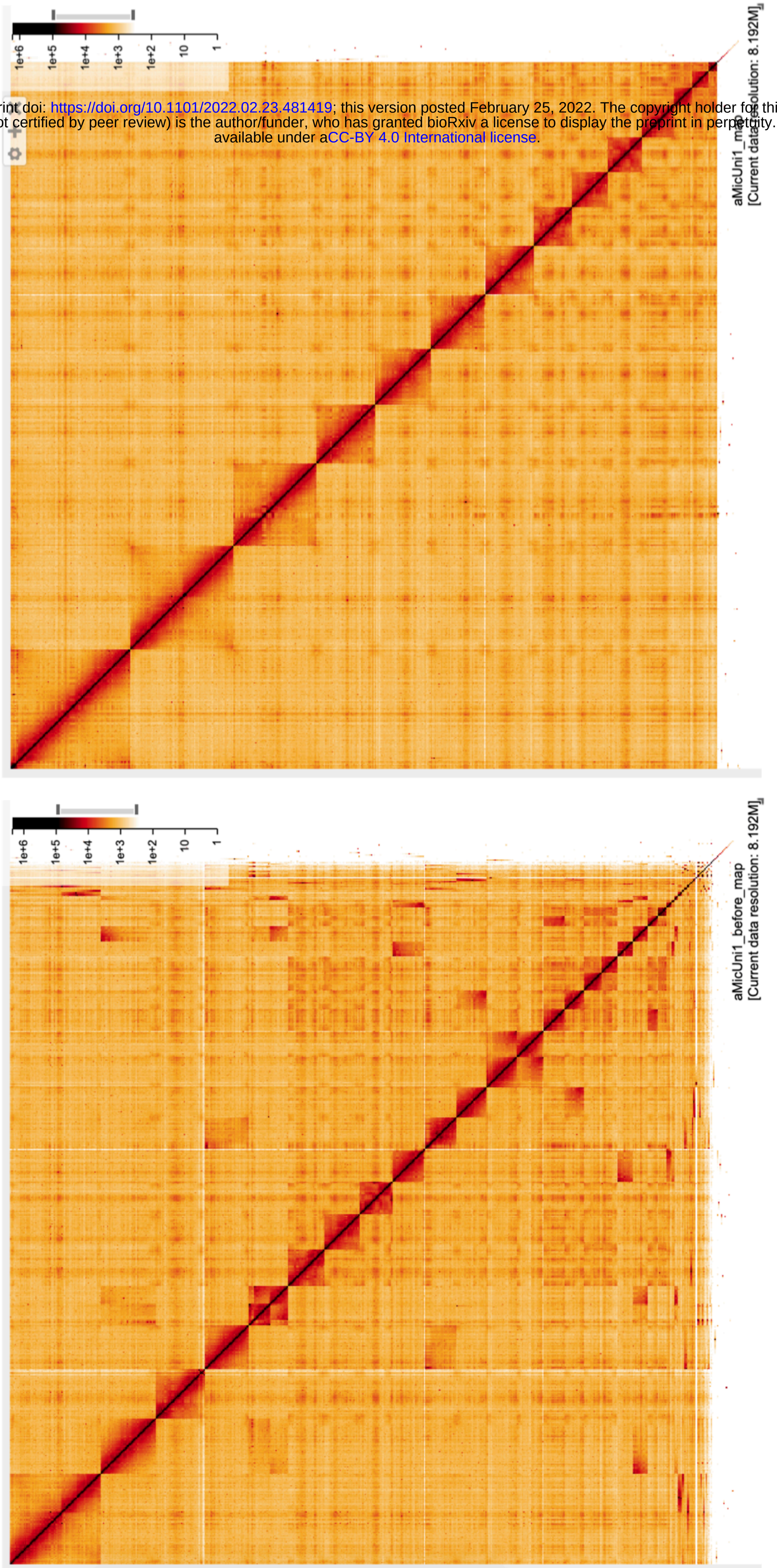
Supplementary Figure S2: Standard Vertebrate Genome Project (VGP) assembly pipeline (vs 1.1-1.6). This diagram is taken from Rhie et al 2021.

Supplementary Figure S3: Phylogeny of vertebrates used in the comparative genomics aspects of the study.

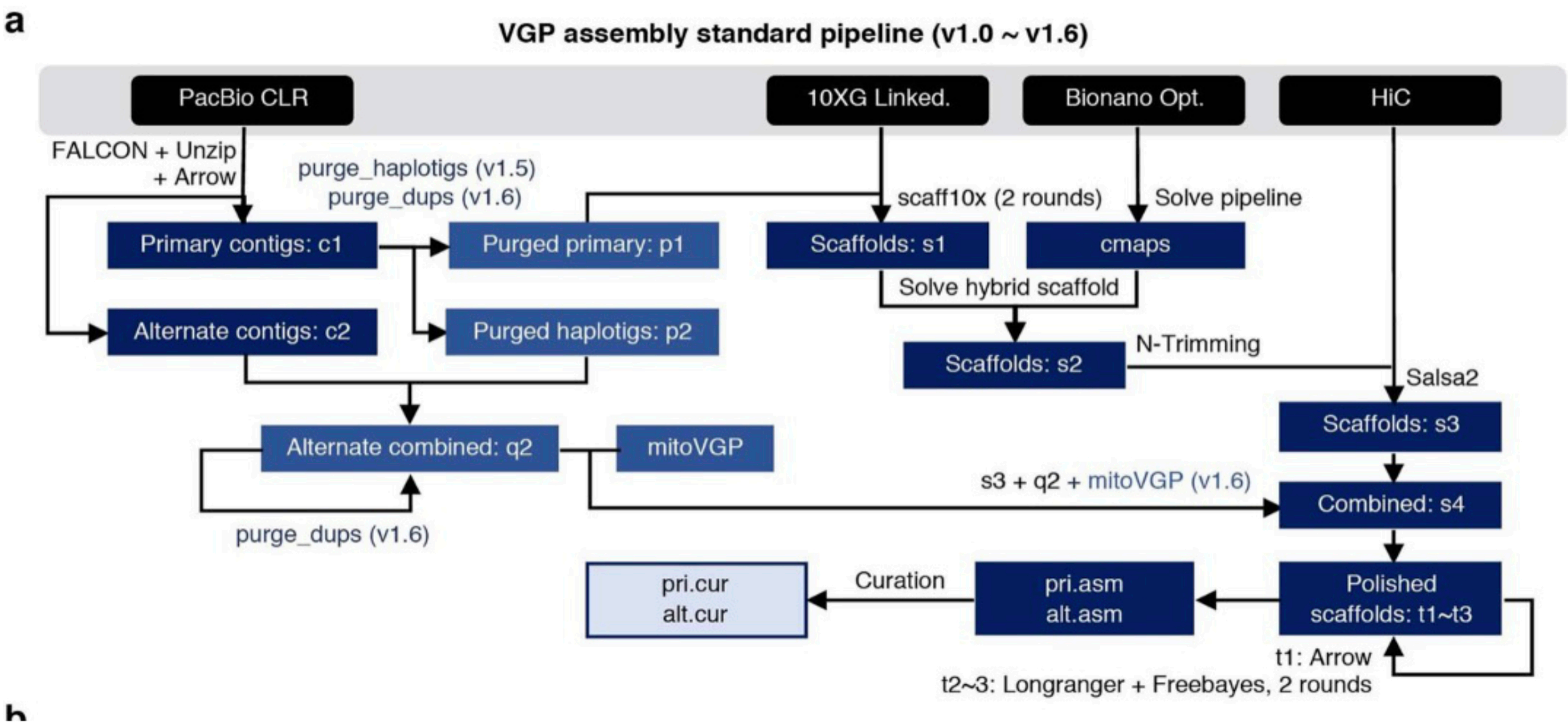
Supplementary Figure S4: Alignment of ZRS enhancer region across a range of vertebrates illustrating the loss of an otherwise well conserved ZRS region in snakes and caecilia.



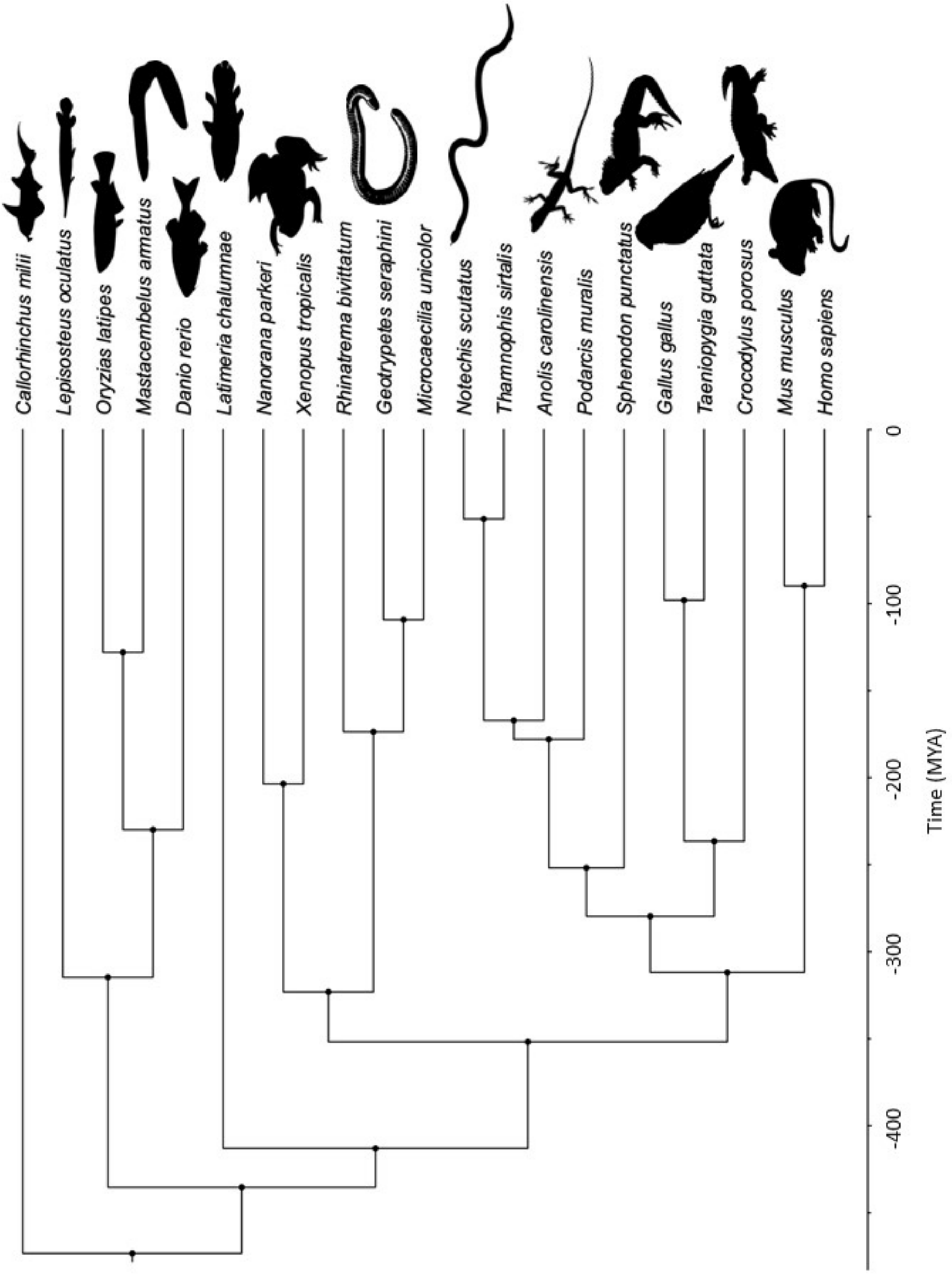
Geotrypetes seraphini Hi-C heatmaps before (left) and after (right) manual curation

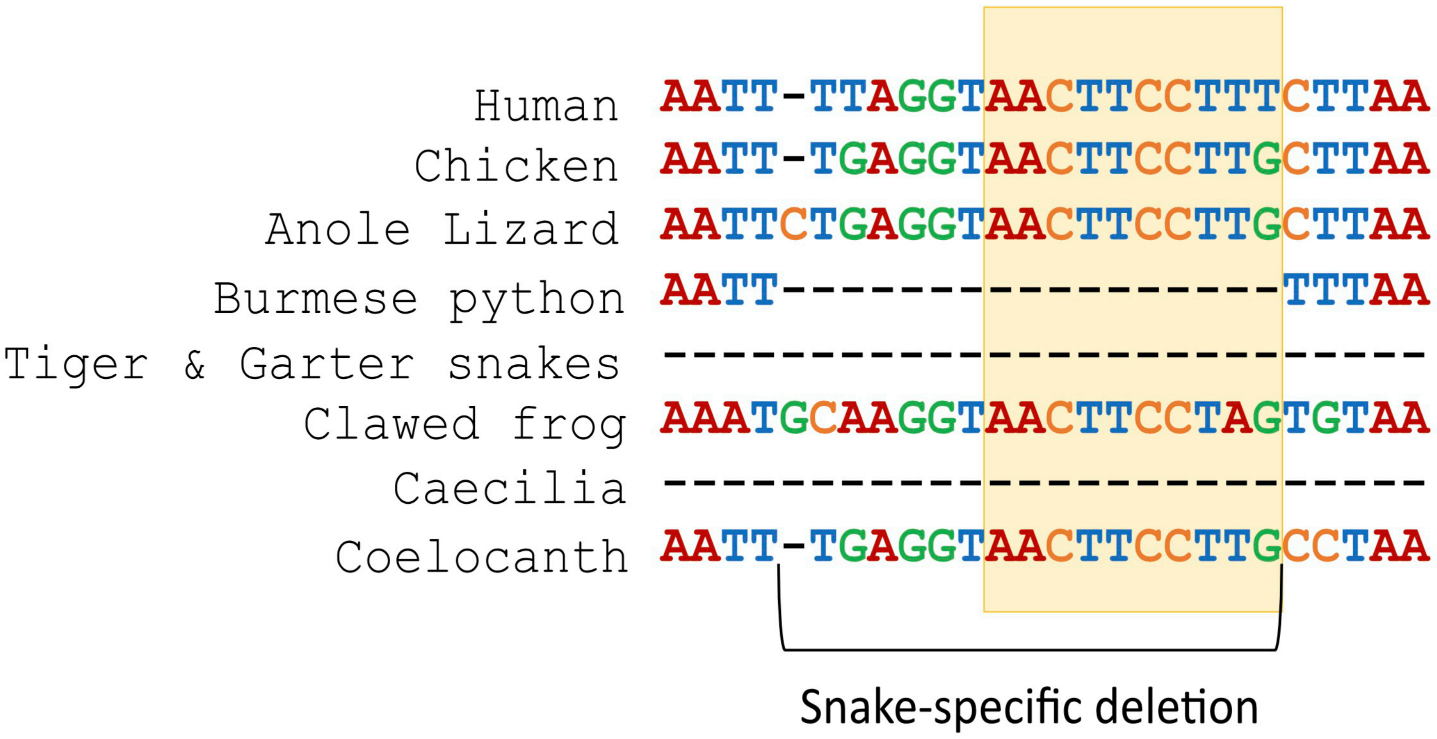


Microcaecilia unicolor Hi-C heatmaps before (left) and after (right) manual curation



Taken from Rhie et al, 2021


Callorhinchus milii
Lepisosteus oculatus
Oryzias latipes
Mastacembelus armatus
Danio rerio
Latimeria chalumnae
Nanorana parkeri
Xenopus tropicalis
Rhinatrema bivittatum
Geotrypetes seraphini
Microcaecilia unicolor
Notechis scutatus
Thamnophis sirtalis
Anolis carolinensis
Podarcis muralis
Sphenodon punctatus
Gallus gallus
Taeniopygia guttata
Crocodylus porosus
Mus musculus
Homo sapiens
0
-100
-200
-300
-400
Time (MYA)


Human AATT-TTAGGTAACTTCCTTTCTTAA
Chicken AATT-TGAGGTAACTTCCTTGCTTAA
Anole Lizard AATTCTGAGGTAACTTCCTTGCTTAA
Burmese python AATT-----------------TTTAA
Tiger & Garter snakes --------------------------
Clawed frog AAATGCAAGGTAACTTCCTAGTGTAA
Caecilia --------------------------
Coelocanth AATT-TGAGGTAACTTCCTTGCCTAA
Snake-specific deletion