The genomic timeline of cichlid fish diversification across continents

Michael Matschiner, Astrid Böhne, Fabrizia Ronco, Walter Salzburger

Cichlid fishes are celebrated for their vast taxonomic, phenotypic, and ecological diversity; however, a central aspect of their evolution — the timeline of their diversification — remains contentious. Here, we generate draft genome assemblies of 14 species representing the global cichlid diversity and integrate these into a new phylogenomic hypothesis of cichlid and teleost evolution that we time-calibrate with 58 re-evaluated fossil constraints and a new Bayesian model accounting for fossil-assignment uncertainty. Our results support cichlid diversification long after the breakup of the supercontinent Gondwana and lay the foundation for precise temporal reconstructions of the exceptional continental cichlid adaptive radiations.

1 Zoological Institute, University of Basel, Basel, Switzerland. 2 Department of Palaeontology and Museum, University of Zurich, Zurich, Switzerland. 3 Centre for Ecological and Evolutionary Synthesis (CEES), Department of Biosciences, University of Oslo, Oslo, Norway. 4 Center for Molecular Biodiversity Research (ZMB), Zoological Research Museum Alexander Koenig, Bonn, Germany. Email: michaelmatschiner@mac.com; walter.salzburger@unibas.ch
wing to their spectacular ecological and morphological diversity and species richness, cichlid fishes have become one of the most important model groups in evolutionary biology and adaptive radiation research\(^1,2\). Despite the great scientific attention that cichlids have received in the last decades, a key aspect of their evolution—the timeline of their diversification and spread to Africa, the Americas, Madagascar, and India—remains controversial\(^3\). Depending on the study, available estimates for the age of the family Cichlidae range from 45 to 160 million years (Myr) and the divergence of the American and African subfamilies (which together include ~99% of cichlid species\(^2\)) has been estimated as recently as 26 million years ago (Ma)\(^4\) or as early as 147 Ma\(^5\). The different age estimates imply contrasting scenarios for the spread of cichlids across continents: although the oldest estimates are compatible with an ancestral cichlid lineage that lived in freshwaters of the former supercontinent Gondwana and diverged by vicariance with its tectonic breakup between 150 and 85 Ma\(^3\), all younger timelines require either long-distance oceanic dispersal events or multiple independent transitions to freshwater from an unknown common marine ancestor (Fig. 1). Because of the requirement of salt-water tolerance or even a marine lifestyle for ancestral lineages of a clade that is confined almost exclusively to freshwater today, the two alternatives to vicariance may appear improbable. On the other hand, salinity tolerance has been observed in some of the most divergent extant cichlid species (reviewed in ref. \(^3\)) and the marine-living convict blenny (\textit{Pholidichthys}) has been identified as the closest living relative to cichlids\(^6\), suggesting that adaptations to marine levels of salinity may have been more common in the early evolution of cichlids.

---

**Fig. 1** Global distribution of cichlid fishes and diversification scenarios.\(^a\) Present-day distribution of cichlid fishes in the Americas (subfamily Cichlinae; blue), Africa, and the Levant (Pseudocrenilabrinae; cyan), Madagascar (Psychochrominae; green), and the Indian subcontinent (Etroplinae; orange). The drawings illustrate the 14 cichlid species used for whole-genome sequencing. Their approximate geographic origins are indicated. \(^b\) Three hypotheses for the phylogeographic history of cichlid fishes. According to the “Gondwanan vicariance” hypothesis, cichlids lived on the Gondwanan landmasses South America, Africa, Madagascar, and India before the separation of these landmasses and diverged as a result of this separation. This would require cichlids to be at least as old as the initial Gondwanan split, i.e., 150 million years. The “Oceanic dispersal” hypothesis posits that cichlids are younger than the separation of Gondwanan landmasses and hence reached their current distributions through long-distance oceanic dispersal. Some molecular studies suggest that this could have occurred around 70 Ma. An alternative hypothesis that is consistent with a young age of cichlids is the “Independent colonization” scenario, according to which cichlids on all four landmasses independently evolved from a common marine ancestor that has since either gone extinct or remained undiscovered. This must have occurred before 45 Ma because the presence of freshwater cichlids by that time is well documented in the fossil record.
The contrasting estimates regarding the timeline of cichlid evolution are due—at least in part—to the use of small phylogenetic datasets dominated by mitochondrial sequences2–9 and to the application of strategies for time calibration that rely exclusively on the cichlid fossil record, without taking into consideration the larger context of teleost evolution, into which the cichlid timeline must be placed10,11. Even if these two issues are addressed, age estimates are still heavily influenced by the often ambiguous assignment of calibration fossils to taxonomic clades, as highlighted in recent studies12–14. For example, when analyzing the same genomic dataset twice with two different fossils that are both currently discussed as potential ambiguous fossil positions, methods have been developed that account for fossil-assignment uncertainty to estimate clade ages in a set of 90 teleosts with a particular focus on cichlid clade ages in a set of 90 teleosts with a particular focus on cichlid clades, of which 7 clades had 2 ambiguous fossil calibrations for 51 clades, of which 7 clades had 2 ambiguous fossil positions, methods have been developed that either infer a fossil’s position during the molecular-clock analysis from scored morphological characters15 or allow multiple positions for one and the same fossil on a fixed tree topology16. However, neither of these methods is suitable for the examination of highly diverse groups of species17. when morphological character matrices are not available and the tree topology is not known a priori.

We here address all three issues that have so far prevented reliable age estimates for cichlid fishes: we (i) provide whole-genome sequencing data for representatives of the global diversity of cichlids, (ii) embed these species into a genome-based phylogeny of teleosts, and (iii) develop and apply a new method to account for uncertain fossil assignments. The resulting timeline based on 91 fish genomes and the fossil record supports the diversification of cichlid fishes long after the breakup of the Gondwanan supercontinent.

Results

Phylogenomic inference of the species-tree topology. We generated draft genome assemblies based on low-coverage Illumina sequencing (7–23x) for 14 cichlid species including 1 species from India, 2 species from Madagascar, 5 from the Americas, and 6 from Africa (Fig. 1 and Supplementary Tables 1 and 2). We then used these whole-genome assemblies, together with a targeted assembly of candidate genes (Supplementary Tables 3–5), to identify 646 single-copy markers with a total alignment length of 127,638 bp for phylogenomic analyses (Supplementary Tables 6–8). Based on these markers, we inferred the species tree for a set of 90 teleost species, including 18 cichlid species, and 1 non-teleost outgroup (Supplementary Figs. 1–4).

Although species-tree estimates produced with the program BEAST218 from concatenated alignments (Supplementary Figs. 1 and 2) agreed well with the current understanding of teleost taxonomy12,13,19, a number of clades that have received unambiguous support from both morphological and molecular datasets (e.g., Acanthomorphpha14,19–22) were not recovered in analyses with the program ASTRAL-I13.23 based on the multispecies coalescent model (Supplementary Figs. 3 and 4). We therefore consider concatenation as the more suitable approach for phylogenomic inference with our dataset. Given the long evolutionary time over which the species in our taxon set diversified (with branch lengths on the order of millions to hundreds of millions of years), the effect of incomplete lineage sorting is likely negligible and the proven statistical inconsistency of concatenation24 due to incomplete lineage sorting is unlikely to affect our conclusions25.

Fossil-based time calibration. To account for ambiguity in fossil assignments, we extended the CladeAge approach for BEAST29,18 so that two fossils can now be specified as potential first records of a clade and weighed according to their relative credibilities. A prior density for the age of the clade is then calculated, taking into account both fossils and their relative credibilities simultaneously (Supplementary Fig. 5). We applied this extended CladeAge approach to time calibrate the teleost species tree with fossil calibrations for 51 clades, of which 7 clades had 2 ambiguous first records (Supplementary Figs. 6 and 7). In six of these seven ambiguous cases, we assigned equal weights to each of two potential first records, naively considering both equally likely to be the true first record of the clade (Supplementary Note 2). The exception to this were the two potential first records of Tetraodontiformes, †P. clarae and †C. guidottii, where we assigned twice the weight to the latter, because we considered it more likely to be the true first record of the clade based on its recent reevaluation12–14,26–28 (Fig. 2).

Our divergence-time estimates are in agreement with the teleost and cichlid fossil records, and pinpoint the age of cichlids at 87.3 Ma (96.9–77.9 Ma; 95% highest posterior density interval), the divergence of the Indian subfamily Etheostomatinae at 76.2 Ma (86.6–66.3 Ma), the separation of the Malagasy subfamily Ptychochrominae at 68.7 Ma (78.0–59.6 Ma), and the divergence between American Cichlinae and African Pseudocrenilabrinae at 62.1 Ma (70.1–54.6 Ma) (Fig. 2), whereby the latter two divergences are close to the Cretaceous-Paleogene boundary, a time of global turmoil29. These estimates are robust to alternative assumptions for the fossilization model used with CladeAge and for the topology of the species tree, and are reduced when the cichlid fossil record is ignored (Table 1). By accounting for uncertainty in fossil assignment, our estimates are also able to resolve the dispute regarding the first record of Tetraodontiformes: as the estimated age of the order is younger than the older of the two potential first records (†P. clarae), our results reject the placement of this fossil within the order.

Discussion

In this study, we generated draft genome assemblies of 14 representative cichlid species and developed a new Bayesian model to account for fossil-assignment uncertainty to estimate clade ages in a set of 90 teleosts with a particular focus on cichlid diversification times, using 58 re-evaluated fossil constraints. Our genomic timeline of cichlid diversification supports the conclusions of earlier studies (reviewed in ref. 3; Supplementary Table 9), which argued against Gondwanan vicariance, given that, e.g., the split between American and African cichlids occurred about 40 Myr after the separation of South America and Africa. On the other hand, our results are unable to distinguish between the two alternative scenarios of post-Gondwanan cichlid diversification: freshwater cichlids could have reached the different landmasses by oceanic dispersal30 or they could have undergone multiple transitions from marine to freshwater to colonize each landmass independently (Fig. 1b).

Both of these scenarios pose questions that our results are unable to answer: if cichlid fishes dispersed from Africa to South America around 62.1 Ma when the Atlantic Ocean was already around 900 km wide30, why is there no evidence of repeated dispersal between Africa and Madagascar across the much narrower Mozambique Channel, which had a width of only 400 km? And if each landmass should have been colonized independently by an unknown marine ancestor, why were most landmasses
apparently colonized only once? Possible explanations for both questions are that perhaps secondary colonizations were unsuccessful due to competition with already-established local cichlid faunas, or that by chance alone the landmasses were colonized just once.

Regardless of these remaining questions concerning the family’s early history, our new timeline of cichlid evolution will be valuable as the basis for the precise temporal reconstructions of the more recent “explosive” adaptive radiations of cichlid fishes that take place in the East African Lakes Tanganyika\(^{31}\), Malawi\(^{32}\), and Victoria\(^{33}\), as well as in numerous other lakes across central Africa and the Neotropics\(^{34–36}\). Owing to their increased precision, these reconstructions may then allow to address important questions about environmental triggers of these radiations such as the roles of lake-level fluctuations\(^{37}\) or ecological opportunity in newly formed lakes\(^{38}\), which could so far not be solved conclusively.

---

**Table 1 Age estimates for selected clades, obtained with different settings and datasets.**

| Setting/Dataset       | Holostei and Teleostei | Cichlidae | Cichlinae and Pseudocrenilabrinae |
|-----------------------|------------------------|-----------|----------------------------------|
| “permissive” gene set | 269.6 (290.0–251.6)    | 76.2 (86.6–66.3) | 62.1 (70.1–54.6) |
| “strict” gene set     | 269.2 (288.5–251.5)    | 76.9 (87.1–66.2) | 62.2 (70.3–54.6) |
| MCMC sampling from prior | 269.6 (290.0–251.5)    | 79.7 (94.9–63.9) | 68.0 (81.3–55.1) |
| Without cichlid calibrations | 268.9 (288.8–251.6)    | 61.2 (77.8–45.5) | 38.7 (50.8–27.1) |
| Doubled net diversification rate | 264.0 (277.5–251.4)    | 73.7 (83.6–64.2) | 60.0 (66.9–53.3) |
| Halved net diversification rate | 277.4 (306.7–251.9)    | 79.7 (90.8–69.4) | 63.5 (72.5–55.0) |
| Doubled fossil sampling rate | 263.5 (279.0–251.4)    | 73.7 (84.2–64.5) | 59.9 (67.2–52.7) |
| Halved fossil sampling rate | 277.9 (302.5–252.5)    | 81.3 (92.3–70.6) | 65.3 (74.1–55.8) |
| (Osteoglossomorpha, Elopomorpha) | 269.4 (289.5–251.5)    | 77.3 (88.1–67.9) | 62.2 (70.3–54.9) |
| (Osteoglossomorpha, Clupeocephala) | 269.2 (289.2–251.5)    | 76.7 (86.1–66.2) | 62.0 (70.1–54.1) |

Unless specified, the “strict” set of genes was used in all analyses. Mean estimates of crown ages in millions of years are given for the three clades, followed by 95% highest posterior density intervals in parentheses. Specified settings in the last two rows indicate monophyly constraints according to alternative relationships among Osteoglossomorpha, Elopomorpha, and Clupeocephala\(^{12, 14}\). The age estimates obtained with the “permissive” gene set correspond to those shown in Fig. 2.
Methods
Species selection for whole-genome sequencing. The species for whole-genome sequencing were selected to cover a wide range of the native cichlid distribution worldwide, including South and Central America, India, Madagascar, Western and Eastern Africa, and to represent all cichlid subfamilies and multiple tribes of the subfamilies Cichlinae and Pseudocrenilabrinae39. Specimens of the species Erythrops canarensis, Paratilapia poloni ‘Andapa’, Pychocromis olaxanthe, Apisto- gramma diptleptis, Astatotilapia burtoni, Amphilophus zalosus, Bujurquina vittata, Andinoacara biseriatus, Heterochromis multimaculatus, Tylochromis polypleis, Benitochromis conjunctus, Pelvicachromis taeniatus, Hemichromis elongatus, and Eita nguti were obtained during field work in Cameroon and Zambia, provided by collaborators or museums, or purchased from the aquarium trade (Supplementary Table 1).

Sequencing. We extracted genomic DNA from fin-clips using the E.Z.N.A. Tissue DNA Kit (Omega Bio-Tek) including an RNase treatment (5 ml, 100 mg/ml for 2 min) and then sheared the DNA on a Covaris E220 (60 W, 2 min) and then sheared the DNA on a Covaris E220 (60 W, 2 min) and then sheared the DNA on a Covaris E220 (60 W, 2 min) and then sheared the DNA on a Covaris E220 (60 W, 2 min) and then sheared the DNA on a Covaris E220 (60 W, 2 min). Library preparation was performed using the Truseq DNA PCR-Free Sample Preparation kit (Low Sample Protocol) for 350 bp insert size. We used a concentration filter with library preparation with optimization and then performed paired-end sequencing of six libraries per lane with a read length of 126 bp on an Illumina HiSeq2500 instrument (v4 chemistry).

Whole-genome assembly. De novo whole-genome assemblies were generated from the Illumina raw sequencing data for each individual following the approach described in Böhne et al.40 and Malmström et al.41 using CeleraAssembler v.8.342 and then sheared the DNA on a Covaris E220 (60 W, 2 min) and then sheared the DNA on a Covaris E220 (60 W, 2 min) and then sheared the DNA on a Covaris E220 (60 W, 2 min) and then sheared the DNA on a Covaris E220 (60 W, 2 min). We assumed the uncorrelated lognormal (UCLN) relaxed molecular clock66 and a model species in the Ensembl database51 to exclude markers with evidence for duplications or deletions. Although previous applications of this strategy were limited to information for 10 teleost species present in the Ensembl database, a massive addition of teleost genomes with previously available genome assemblies for 76 teleost species (including four further cichlid species) and one non-teleost outgroup, for a total of 91 assemblies (Supplementary Tables 6 and 7). These species were selected to represent all major lineages within teleosts with increased sampling density for cichlids closely related to cichlid fishes (e.g., orthologs of other cichlids and other groups within the series Ovatentaria). As a non-teleost outgroup, we included the spotted barb (Lepisosteus oculatus), a member of Holostei, the sister group of Teleostei67. We chose this sampling scheme, as it allowed us to use a wide range of fossil occurrences, including the earliest records of teleosts and their sister group, to calibrate the origin and the timeline of teleost evolution.

Marker selection. As teleost fishes began to diverge over 200 Ma46,47 and their genomes have undergone duplications and frequent rearrangements48, it is difficult to reliably determine the orthology of most genomic regions across divergent teleost species. We therefore focused on conserved coding genes as the most suitable type of markers for phylogenomic inference, using a strategy that has already been applied successfully in several studies of teleost diversity times14,16,20,27. This strategy makes use of the information on gene relationships among teleost model genes in the Ensembl database43 to exclude markers with evidence for duplications or deletions. Although previous applications of this strategy were limited to information for 10 teleost species present in the Ensembl database, a massive addition of teleost genomes with previously available genome assemblies for 76 teleost species (including four further cichlid species) and one non-teleost outgroup, for a total of 91 assemblies (Supplementary Tables 6 and 7). These species were selected to represent all major lineages within teleosts with increased sampling density for cichlids closely related to cichlid fishes (e.g., orthologs of other cichlids and other groups within the series Ovatentaria). As a non-teleost outgroup, we included the spotted barb (Lepisosteus oculatus), a member of Holostei, the sister group of Teleostei67. We chose this sampling scheme, as it allowed us to use a wide range of fossil occurrences, including the earliest records of teleosts and their sister group, to calibrate the origin and the timeline of teleost evolution.

Ortholog identification and filtering. Similar to the marker selection procedure, the identification and filtering of sequences orthologous to the selected markers followed the workflow established in Malmström et al.49 and refined in Musilova et al.14. To compile a first set of candidate orthologs, we used each of the selected 5869 zebrafish exons as query and each of the 91 genome assemblies as subjects in TBLASTN searches. Per combination of exon query and assembly subject, we calculated the best hit as a candidate ortholog if its bitscore was above the exon-specific bitscore threshold during the marker selection. The TBLASTN searches produced a total of 488,171 sequences for the 5869 exons. Per exon, we then aligned all candidate orthologs using MAFFT v.7.3095, while ensuring the integrity of codon triplets. Exon alignments were then filtered according to several criteria, aiming to exclude potentially remaining paralogs and to select the most suitable alignments for phylogenetic divergence-time estimation: (1) we excluded sequences with TBLASTN bitscore values lower than 90% of the highest bitscore value achieved by any of the ingroup sequences; this removed 130,321 out of 488,171 sequences across the 5869 exon alignments. (2) We calculated dN/dS between each ingroup subject and the outgroup sequence and used the PAML 4.8.1 package49, and removed sequences with dN/dS ratios > 0.25, as high dN/dS ratios can indicate positive selection, reading-frame shifts, or intronic regions. This filter removed 2470 sequences overall. (3) We excluded all exon alignments that had more than 10 missing sequences; this removed 4178 of the remaining alignments. (4) We then excluded all alignments containing at least one site with more than 20% missing data or a smoothed entropy-like score above 0.5, determined with the program BMGE v.11.02. This filter removed 6035 codons and thus 18,105 bp out of 373,545 bp. (5) We excluded exon alignments shorter than 130 bp, this filter removed 71 of the 1690 remaining alignments. (6) As high GC-content variation has been shown to potentially misguide phylogenetic inference68, we excluded all exons with an among-sequence GC content in GC content above 0.04; this removed 247 of the remaining 1619 exon alignments. (7) We excluded the exons of all genes if these genes did not have at least three exons remaining in the dataset or if their exons were over 100,000 bp apart on the zebrafish genome; we removed 1723 of the remaining 3727 exon alignments. (8) We removed all exons with exon trees that were significantly discordant to the trees of other exons of the same gene, which could potentially result from paralogy. As in Malmström et al.49 and Musilova et al.14, these concordance tests were performed with the program Concaterpillar v.1.7-2.249, which internally uses RAxML v.7.2.849 and the generalized time-reversible (GTR) model of sequence evolution49 for maximumlikelihood tree inference. Following the concordance tests, we concatenated, per gene, the alignments of all exons with concordant exon trees. Genes that did not have at least three exons with concordant exon trees were removed. (9) The concatenated exon alignments of all remaining 161 genes were visually checked to avoid potential alignment errors69 and two genes were excluded due to potential total misalignment.

Finally, we quantified the substitution rate and its among-species variation for each gene in separate molecular-clock analyses with the Bayesian software BEAST 2 v2.5.038, and we selected two nested sets of genes according to different threshold for these molecular clock parameters. The models used in these BEAST 2 analyses assumed the uncorrelated lognormal (UCLN) relaxed molecular clock46 and a pure-birth speciation process47, and the mBModelTest package38 for BEAST 2 was
employed to average over nucleotide substitution models. In each BEAST 2 analysis, the Markov-chain Monte Carlo (MCMC) process was run for 50 million iterations, which provided sufficient effective sample sizes (ESS) for all model parameters of at least 200 for all but 14 genes (ESS values reached at least 100 in all but 5 genes). As low ESS values can result from conflicting phylogenetic signals within genes or low information content of the alignment, we considered these ESS values in our selection of genes for further phylogenomic analyses. Thus, we selected a “permissive” set of genes that included all genes with a minimum ESS value above 200, a substitution rate below 1.6 x 10^{-9} per year and site, and a coefficient of rate variation below 0.7; this set included 147 genes with a total alignment length of 127,638 bp. In addition, we selected a “strict” set of genes as a subset of the “permissive” set including all genes with a minimum ESS value above 100, a substitution rate below 1.4 x 10^{-9} per year and site, and a coefficient of rate variation below 0.6; this set included 77 genes with a total alignment length of 62,776 bp (Supplementary Table 8). The two sets complemented each other as the “permissive” set was expected to be more phylogenetically informative due to its larger size and higher mean substitution rate, whereas the “strict” set was expected to contain less homoplasies and evolve in a more clock-like manner, both of which could lead to more accurate age estimates.

**Species-tree inference.** We performed species-tree inference with both the “permissive” and the “strict” set of genes, and applied both the multi-species coalescent model and concatenation. Analyses with the multi-species coalescent model were conducted with the software ASTRAL-III v.5.6.23, using maximum-clade credibility summary trees of the per-genome BEAST 2 analyses as input. The concatenated alignments of the “permissive” and “strict” gene sets were separately analyzed with BEAST 2, applying the birth–death tree prior49, the UCLN relaxed clock model, and independent GTR site models with γ-distributed rate variation for eight (“permissive”) or five (“strict”) partitions. One-over-x priors distributions were placed on the mean mutation rates of each partition and an exponential prior distribution (with a mean of 0.5) was used for the SD of among-branch rate variation. The partitions were selected from the set of all first and second codon positions, using the cluster algorithm of the software PartitionFinder v.223,24, with equal weights for all model parameters and a minimum partition size of 5000 sites.

Third codon positions were excluded from this set of BEAST 2 analyses to reduce both the computational demand and the degree of saturation in the alignment. As these BEAST 2 analyses aimed to infer the species-tree topology rather than its divergence times, they were time-calibrated only by an age constraint on the root node, which was assigned a set to 300 Ma with a SD of 0.1 Myr. To facilitate MCMC convergence, a single monophyley constraint was placed to group Syngnatharia and Pelagia. The sister-group relationship of these two clades is overwhelmingly supported by molecular data2,23,19, but is difficult to infer in molecular-clock analyses due to highly divergent substitution rates of the two lineages. For each of the “permissive” and “strict” gene sets, we performed 10 replicate BEAST 2 analyses, each with 300 million MCMC iterations. Convergence of the MCMC was confirmed by ESS values of at least 200 (“permissive”) or 400 (“strict”) for all model parameters.

**Accounting for fossil-assignment uncertainty.** The CladeAge add-on package for BEAST 24 replaces the specification of lognormal, normal, uniform, or other types of distributions for prior densities with automatically calculated prior densities that are shaped according to expectation for clad age under various assumptions for diversification parameters and the fossil sampling rate. This calibration framework is well suited to account for uncertainty in fossil assignment. With a single unambiguous calibration framework is well suited to account for uncertainty in fossil assignment. With a single unambiguous calibration density and the MCMC trace revealed that the chain switched frequently between the two peaks; thus, we found no signs of poor convergence (Supplementary Fig. 5).

**Phylogenetic divergence-time estimation.** Teleost divergence times were estimated based on the CladeAge approach in which calibration priors are calculated from estimates of fossil age, diversification rates, and the fossil sampling rate. The model underlying this type of divergence-time estimation assumes that prior distributions are defined for all clades that fulfill the following three conditions: (1) the clade must be represented in the fossil record, (2) the clade must be morphologically recognizable that fossils can be assigned directly to it, not only indirectly through assignment to a subclade, and (3) all potential sister lineages of this clade must be present in the phylogeny so that the origin of this clade is guaranteed to be represented by a node in the phylogeny. For 51 clades that matched the criteria for CladeAge calibrations, we identified the earliest fossil occurrences, determined their geologic ages, and the absolute ages of the stage (or the stratigraphic age) used to define age constraints with CladeAge (Supplementary Note 2). The first occurrences of seven clades were found to be ambiguous with two fossils in each case that could potentially represent the clades’ earliest records. In these cases, both potential first occurrences were used in the analyses, with weights as specified in Supplementary Note 2. We assumed the same estimates for the teleost fossil sampling rate (0.0066–0.01806 per lineage per Myr73), their net diversification rate (0.041–0.081 per lineage per Myr73), and their turnover (0.0011–0.037 per lineage per Myr73) as in Matschiner et al.39. To fix the tree topology to that of the species tree inferred from the concatenated “permissive” gene alignments (see above), we used this species tree as the starting tree and disabled all topology operators. As in the earlier analyses of the species-tree topology based on concatenation, we performed phylogenetic divergence-time estimations separately with both the “permissive” and the “strict” set of gene alignments, and we applied the same partitioning schemes as in these earlier analyses. The settings for the assumed substitution model (the GTR model with γ-distributed rate variation) and the tree prior (the birth–death tree prior) were also identical to the earlier analyses of the species-tree topology. We again performed ten replicate BEAST 2 analyses for both the “permissive” and the “strict” gene set, in each case with 100 million MCMC iterations. These analyses produced ESS values > 200 (“permissive”) or 1000 (“strict”) for all model parameters.

We assessed the robustness of our divergence-time estimates with a range of analyses that were identical to those for the “strict” dataset except that (1) MCMC sampling was done without data, only from the prior distributions; (2) all cichlid fossils were excluded; (3) the fossil sampling rate assumed for CladeAge calibrations was doubled or halved; (4) the net diversification rate assumed for calibrations was doubled or halved; and (5) the sequence alignment was used to define age constraints with CladeAge (Supplementary Note 2) or to Osteoglossomorpha14,24 or to Clupeocephala12,19.

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

**Data availability** Data generated for this study are available from NCBI under the BioProject accession number PRJNAS502935. Previously available datasets used in this study are either hosted at the Ensembl (Ensemb.org), NCBI (ncbi.nlm.nih.gov), or EBI (ebi.ac.uk) databases, or deposited on dataaday.org, figshare.com, parrot.genomics.cn, surfdrive.surf.nl, cichlid.gurdon.cam.ac.uk, eflshgenomics.integrativebiology.msu.edu, or cresskolab.oregon.edu (see Supplementary Table 7 for details). Sequence alignments used for phylogenomic inference are available from http://evoinformatics.eu/continental.htm. Figure 2 and another
Supplementary Figs. 1–4, 6, and 7 have associated raw data available from http://evoinformatics.eu/continental.htm.

Code availability

Code for computational analyses is available from Github (http://github.com/mmatschiner/continental).

Received: 26 February 2020; Accepted: 15 July 2020; Published online: 18 November 2020

References

1. Brawand, D. et al. The genomic substrate for adaptive radiation in African cichlid fish. Nature 513, 375–381 (2014).
2. Salzburger, W. Understanding explosive diversification through cichlid fish genomics. Nat. Rev. Genet. 19, 705–717 (2018).
3. Matschiner, M. Gondwanan vicariance or trans-Atlantic dispersal of cichlid fishes: a review of the molecular evidence. Hydrobiologia 832, 9–37 (2019).
4. Near, T. J. et al. Phylogeny and tempo of diversification in the superradiation of spiny-rayed fishes. Proc. Natl Acad. Sci. USA 110, 12738–12743 (2013).
5. López-Fernández, H., Arbour, J. H., Winemiller, K. O. & Honeycutt, R. L. Testing for ancient adaptive radiations in neotropical cichlid fishes. Evolution 67, 1321–1337 (2013).
6. Etyan, R. I. et al. Interfering acanthomorph teleost phylogeny using anchored hybrid enrichment. BMC Evol. Biol. 15, 113 (2015).
7. Azuma, Y., Kumazawa, Y., Miyaji, M., Mabuchi, K. & Nishida, M. Mitogenomic evaluation of the historical biogeography of cichlids toward resolving dating of teleostean divergences. BMC Evol. Biol. 8, 215 (2008).
8. McMahan, C. D., Chakraborty, P., Sparks, J. S., Smith, W. L. & Davis, M. P. Temporal patterns of diversification across global cichlid biodiversity (Acanthomorpha: Cichlidae). PLoS ONE 8, e71162 (2013).
9. Matschiner, M. et al. Bayesian phylogenetic estimation of clade ages supports multiple radiations interconnected by gene flow in African cichlid fishes. Nature https://doi.org/10.1038/s41586-020-2930-4 (2020).
10. Malinsky, M. et al. Whole-genome sequences of Malawi cichlids reveal multiple radiations interconnected by gene flow. Nat. Ecol. Evol. 4, 1940–1955 (2020).
11. Meier, J. I. et al. Ancient hybridization fuels rapid cichlid fish adaptive radiations. Nat. Commun. 8, 14363 (2017).
12. Wagner, C. E., Harmon, L. I. & Seehausen, O. Ecological opportunity and sexual selection together predict adaptive radiation. Nature 487, 366–369 (2012).
13. Elmer, K. R. et al. Parallel evolution of Nicaraguan crater lake cichlid fishes via non-parallel routes. Nat. Commun. 5, 1–8 (2014).
14. Martin, C. H. et al. Complex histories of repeated gene flow in Cameroon crater lake cichlids cast doubt on one of the clearest examples of sympatric speciation. Evolution 69, 1406–1422 (2015).
15. Ivory, S. J. et al. Environmental change explains cichlid adaptive radiation at Lake Malawi over the past 1.2 million years. Proc. Natl Acad. Sci. USA 113, 19195–19200 (2016).
16. Böhne, A. et al. Repeated evolution versus common ancestry: sex chromosome evolution in the haptophycine cichlid Pseudocrenilabrus phillander. Genome Biol. Evol. 11, 439–458 (2019).
17. Malmström, M. et al. Whole genome sequencing data and de novo draft assemblies for 66 teleost species. Sci. Data 4, 160132 (2017).
18. Myers, E. W. et al. A whole-genome assembly of Drosophila. Science 287, 2196–2204 (2000).
19. Magoc, T. & Salzberg, S. L. FLASH: fast length adjustment of short reads to improve genome assemblies. Bioinformatics 27, 2957–2963 (2011).
20. Gurevich, A., Saveliev, V., Vyahhi, N. & Tesler, G. QUAST: quality assessment tool for genome assemblies. Bioinformatics 29, 1072–1075 (2013).
21. Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics 31, 3210–3212 (2015).
22. Arratia, G. Morphology, taxonomy, and phylogeny of Triassic pholidophorid fishes (Actinopterygii) based on transcriptomic and genomic data. Proc. Natl Acad. Sci. USA 5, 201719358 (2018).
23. Alfaro, M. et al. Explosive diversification of marine fishes at the Cretaceous–Palaeogene boundary. Nat. Ecol. Evol. 2, 688–696 (2018).
24. Musilova, Z. et al. Evolution using multiple distinct rod opsins in deep-sea fishes. Science 354, 588–592 (2016).
25. Gavryushkina, A. et al. Bayesian total-evidence dating reveals the recent crown radiation of penguins. Syst. Biol. 66, 57–73 (2017).
26. Guindon, S. Accounting for calibration uncertainty: Bayesian molecular dating as a “doubly intractable” problem. Syst. Biol. 67, 651–661 (2018).
27. Matschiner, M. Selective sampling of species and fossils influences age estimates under the fossilized birth/death model. Front. Genet. https://doi.org/10.3389/fgene.2019.00164 (2019).
28. Bouckaert, R. R. et al. BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis. PLoS Comput. Biol. 15, e1006650 (2019).
29. Betancur-R. et al. Phylogenetic classification of bony fishes. BMC Evol. Biol. 17, 162 (2017).
30. Johnson, G. D. & Patterson, C. Percomorph phylogeny: a survey of Acanthomorphs and a new proposal. Bull. Mar. Sci. 52, 554–626 (1993).
31. Wiley, E. O. & Johnson, G. D. A teleost classification based on monophyletic groups. In Origin and Phylogenetic Interrelationships of Teleosts 123–182 (Verlag Dr. Friedrich Pfeil, München, Germany, 2010).
32. Chen, W.-J. et al. New insights on early evolution of spiny-rayed fishes (Teleostei: Acanthomorpha). Front. Mar. Sci. 1, 53 (2014).
33. Zhang, C., Rabbee, M., Sayyari, E. & Mirarab, S. ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees. BMC Bioinformatics 19, 153 (2018).
34. Nordsiek, S. & Steel, M. Likelihood-based tree reconstruction on a concatenation of aligned sequence data sets can be statistically inconsistent. Theor. Popul. Biol. 100, 56–62 (2014).
35. Bryant, D. & Hahn, M. W. The concatenation question. In Phylogenetics in the Genomic Era (eds Scornavacca, C., Deluc, F. & Galtier, N.) 3.4.1–3.4.23 (No commercial publisher, 2020).
Acknowledgements

We thank Marta Barluenga, Adrian Indermaur, and Zuzana Musilová for providing samples, Remco Bouckaert for support with CladeAge development, and Alex Viertler for the cichlid drawings. Funding was provided by the Norwegian Research Council (FRIPRO project 275869 to M.M.), the Swiss National Science Foundation (SNSF grant Ambizione PZ00P3_161462 to A.B.), and the European Research Council (ERC Consolidator Grant Number 617585 ‘CICHLID~X’ to W.S.). Calculations for genome assemblies were done at the sciCORE (http://scicore.unibas.ch) scientific computing center at the University of Basel; all other computational analyses were performed on the Abel supercomputing cluster (Norwegian metacenter for High Performance Computing and the University of Oslo) operated by the Research Computing Services group at USIT, the University of Oslo IT department.

Author contributions

M.M. and W.S. conceived this study. M.M. performed phylogenomic analyses and developed software. A.B. and F.R. extracted DNA and prepared genome-sequencing libraries. A.B. assembled genomes. M.M. and W.S. drafted the manuscript and all authors commented on it.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41467-020-17827-9.

Correspondence and requests for materials should be addressed to M.M. or W.S.

Peer review information Nature Communications thanks Julia Day, Mario dos Reis Barros, and the other, anonymous, reviewer for their contribution to the peer review of this work.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher’s note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s) 2020