Molecular Phylogenetics and Evolution 83 (2015) 56–71

Contents lists available at ScienceDirect

Molecular Phylogenetics and Evolution

journal homepage: www.elsevier.com/locate/ympev

A tribal level phylogeny of Lake Tanganyika cichlid fishes based on a genomic multi-marker approach

Britta S. Meyer [a,b,*], Michael Matschiner [a,c], Walter Salzburger [a,c,*]

[a] Zoological Institute, University of Basel, Vesalgasse 1, 4051 Basel, Switzerland
[b] Evolutionary Ecology of Marine Fishes, GEOMAR Helmholtz Centre for Ocean Research Kiel, Düsternbrooker Weg 20, 24105 Kiel, Germany
[c] Centre for Ecological and Evolutionary Synthesis (CEES), Department of Biosciences, University of Oslo, Oslo, Norway

ARTICLE INFO

Article history:
Received 25 November 2013
Revised 5 October 2014
Accepted 8 October 2014
Available online 26 November 2014

Keywords:
Adaptive radiation
Cichlidae
454 amplicon sequencing
Hybridization
Incomplete lineage sorting

ABSTRACT

The species-flocks of cichlid fishes in the East African Great Lakes Victoria, Malawi and Tanganyika constitute the most diverse extant adaptive radiations in vertebrates. Lake Tanganyika, the oldest of the lakes, harbors the morphologically and genetically most diverse assemblage of cichlids and contains the highest number of endemic cichlid genera of all African lakes. Based on morphological grounds, the Tanganyikan cichlid species have been grouped into 12–16 distinct lineages, so-called tribes. While the monophyly of most of the tribes is well established, the phylogenetic relationships among the tribes remain largely elusive. Here, we present a new tribal level phylogenetic hypothesis for the cichlid fishes of Lake Tanganyika that is based on the so far largest set of nuclear markers and a total alignment length of close to 18 kb. Using next-generation amplicon sequencing with the 454 pyrosequencing technology, we compiled a dataset consisting of 42 nuclear loci in 45 East African cichlid species, which we subjected to maximum likelihood and Bayesian inference phylogenetic analyses. We analyzed the entire concatenated dataset and each marker individually, and performed a Bayesian concordance analysis and gene tree discordance tests. Overall, we find strong support for a position of the Oreochromini, Boulengerochromini, Bathybatini and Trematocarini outside of a clade combining the substrate spawning Lamprologini and the mouthbrooding tribes of the 'H-lineage', which are both strongly supported to be monophyletic. The Eretmodini are firmly placed within the 'H-lineage', as sister-group to the most species-rich tribe of cichlids, the Haplochromini. The phylogenetic relationships at the base of the 'H-lineage' received less support, which is likely due to high speciation rates in the early phase of the radiation. Discordance among gene trees and marker sets further suggests the occurrence of past hybridization and/or incomplete lineage sorting in the cichlid fishes of Lake Tanganyika.

1. Introduction

The species-flocks of cichlid fishes in the East African Great Lakes Victoria, Malawi and Tanganyika (LT) represent the most species-rich adaptive radiations known in vertebrates (see e.g. Kocher, 2004; Salzburger, 2009; Seehausen, 2006). Several hundred of endemic cichlid species have evolved in each of these lakes in only the last few million to several thousand of years (see e.g. Genner et al., 2007; Kocher, 2004; Salzburger, 2009; Salzburger and Meyer, 2004; Snoeks, 2000; Turner et al., 2001; Verheyen et al., 2003). Because of their taxonomic diversity, their ecological and morphological disparity and the high proportion of endemism, East African cichlid fishes are a prime model system in evolutionary biology (reviewed in: Kocher, 2004; Salzburger, 2009; Santos and Salzburger, 2012).

With a maximum estimated age of 9–12 million years (my), LT is the oldest lake in Africa (Cohen et al., 1997; Salzburger et al., 2014) and contains the genetically, morphologically and ecologically most diverse group of cichlid fishes counting ca. 200 species in more than 50 genera (Koblmüller et al., 2008b; Salzburger et al., 2002a; Snoeks, 2000). Based on morphological grounds, Poll (1986) grouped the LT cichlid species into 12 tribes (a taxonomic rank between subfamily and genus): Bathybatini, Cyprichromini, Ectodini, Eretmodini, Haplochromini, Lamprologini, Limnochromini, Perissodini, Tilapiini, Trematocarini, Tropheini,

* Corresponding authors at: Evolutionary Ecology of Marine Fishes, GEOMAR Helmholtz Centre for Ocean Research Kiel, Düsternbrooker Weg 20, 24105 Kiel, Germany (B.S. Meyer). Zoological Institute, University of Basel, Vesalgasse 1, 4051 Basel, Switzerland (W. Salzburger).

E-mail addresses: britta-meyer@gmx.de (B.S. Meyer), walter.salzburger@unibas.ch (W. Salzburger).

http://dx.doi.org/10.1016/j.ympev.2014.10.009
1055-7903/

and Tylochromini. Takahashi (2003) revised Poll's tribal assignment and suggested to (i) taking *Boulengerochromis microlepis* out of the Tilapiini into its own tribe, Boulengerochromini, leaving behind *Oreochromis tanganicae* as the only representative of the Tilapiini in LT; (ii) splitting the Limnochromini into Limnochromini *sensu stricto*, Benthochromini and Greenwoodochromini; (iii) establishing a separate tribe, Cyphotilapiini, for *Cyphotilapia frontosa* and *C. gibberosa*; (iv) moving '*Ctenochromis*' *benthicola* into its own tribe; and (v) putting the species of the Trematocarini into the Bathybatini. Only some of these revisions are backed up by molecular data, such as the establishment of the new tribes Benthochromini, Boulengerochromini, and Cyphotilapiini (Koblmüller et al., 2008b; Muschick et al., 2012; Salzburger et al., 2002a). *Greenwoodochromis*, on the other hand, is clearly nested within the Limnochromini in molecular phylogenies (Duftner et al., 2005; Muschick et al., 2012; Kirchberger et al., 2014), and should hence remain within the Limnochromini; the Trematocarini consistently form a separate lineage outside the Bathybatini (see e.g. Koblmüller et al., 2005; Muschick et al., 2012) and should remain in their own tribe (note that Koblmüller et al. (2008b) suggested splitting the Bathybatini into Bathybatini *sensu stricto* and Hemibatini); and '*Ctenochromis*' *benthicola* has recently been identified as member of the Cyphotilapiini (Muschick et al., 2012). Finally, the Tropheini were consistently found to be nested within the Haplochromini (Salzburger et al., 2005, 2002a; see also below) and should, hence, not be considered as separate tribe but as part of the Haplochromini.

Not all of the cichlid tribes occurring in LT are endemic to this lake, though, and four tribes show a distribution range that exceeds the LT basin by far. The Tylochromini have their center of divergence in West Africa (Stiassny, 1990), and the only LT species, *T. polylepis*, is likely to have invaded LT only recently (Koch et al., 2007). The same might be true for *O. tanganicae*, the only native representative of the widely distributed Tilapiini in LT (Klett and Meyer, 2002). Note that the Tilapiini were recently taxonomically revised and that the genus *Oreochromis* has been placed into a new tribe, namely the Oreochromini (Dunz and Schliewen, 2013). The Lamprologini, the most species-rich tribe of cichlids in LT, contain a few species that have secondarily colonized the Congo and Malagarasi River systems (Salzburger et al., 2002a; Schelly et al., 2003; Schelly and Stiassny, 2004; Sturmbauer et al., 2010). The Haplochromini (including the Tropheini) represent the most species-rich tribe of cichlids overall, and are distributed across large parts of Africa, where they have seeded various radiations including the ones of Lake Malawi and the Lake Victoria Region (Koblmüller et al., 2008a; Salzburger et al., 2005; Schwarzer et al., 2012; Verheyen et al., 2003; Wagner et al., 2012). The LT cichlid fishes thus show faunal affinities across a large geographical range to both older cichlid lineages such as the Tylochromini and Tilapiini/Oreochromini and younger ones such as the Haplochromini.

The phylogenetic relationships among East African cichlid tribes has been the subject of various studies over the past two decades, yet remain enigmatic (reviewed in: Koblmüller et al., 2008b). The first comprehensive phylogenetic study of LT's cichlid fishes using molecular information dates back to the early 1990s, when Nishida (1991) used allozyme data to examine the relationships among tribes. He established the so-called 'H-lineage' consisting of the tribes Cyprichromini, Ectodini, Eretmodini, Haplochromini/Tropheini (which he already found to be monophyletic), Limnochromini, and Perissodini as sister-group to the Lamprologini; the Bathybatini, Trematocarini plus *Boulengerochromis microlepis*, *Oreochromis tanganicae*, and *Tylochromis polylepis* were placed outside of a clade formed by the 'H-lineage' and Lamprologini. Yet, the relative position of the 'H-lineage' tribes differed depending on the algorithms used (UPGMA and neighbour-joining; NJ) (Fig. 1a). Sturmbauer and Meyer (1993) used two mitochondrial (mt) DNA markers (cytochrome *b* and control region) and suggested, based on phylogenetic analyses with NJ and maximum parsimony (MP), a sister-group relationship between the Cyprichromini and the Ectodini and between the Eretmodini and the Haplochromini (Fig. 1b). Kocher et al. (1995) established the mitochondrial NADH dehydrogenase subunit 2 (ND2) gene as marker for phylogenetic analyses in cichlid fishes and provided the most inclusive phylogenetic hypothesis for LT cichlids so far. In their MP and NJ phylogenies, the Bathybatini, the Tylochromini, *B. microlepis* and *O. tanganicae* formed a clade, and the Eretmodini were placed outside the 'H-lineage', as sister-group to the Lamprologini (Fig. 1c). The Cyprichromini were resolved as the sister-group to all remaining 'H-lineage' taxa (i.e. without the Eretmodini). Using three mitochondrial markers (control region, cytochrome *b*, ND2) and NJ, MP and maximum-likelihood (ML) phylogenetic analyses, Salzburger et al. (2002a) confirmed the position of *B. microlepis*, the Bathybatini and the Trematocarini outside all other tribes occurring in Lake Tanganyika, with the exception of the Tylochromini, and the Eretmodini were placed as sister-group to the Lamprologini and the remaining 'H-lineage' tribes (Fig. 1d). Within the 'H-lineage', the Ectodini appeared as the sister to the remaining taxa. This study was also the first to establish phylogenetic affinities between the LT cichlid fishes and the riverine genus *Orthochromis* (not shown in Fig. 1d; see also Salzburger et al., 2005). Clabaut et al. (2005) combined sequences of the mitochondrial ND2 gene and the nuclear recombinase activating gene (*rag*) and applied ML and Bayesian inference (BI). They placed the Eretmodini as sister-group to the Lamprologini and established the 'C-lineage', i.e. the 'H-lineage' of Nishida (1991) but without the Eretmodini. Within this 'C-lineage', the Limnochromini plus *C. frontosa* appeared as the sister-group to the Perissodini, the Ectodini, the Cyprichromini and the Haplochromini (Fig. 1e). Day et al. (2008) provided one of the most comprehensive datasets to date (cytochrome *b*, ND2) including 157 taxa. Their ML and BI phylogenies supported the existence of the 'C-lineage' by placing the Eretmodini as sister-group to the Lamprologini. In their analyses, a clade formed by the Ectodini and Cyprichromini was placed as the sister-group of the remaining 'C-lineage' taxa (Fig. 1f). In the ML phylogeny of Muschick et al. (2012), who used the mitochondrial ND2 gene and two nuclear markers (*ednrb1*, *phpt1*), the Eretmodini were placed as sister group to the Lamprologini and the 'C-lineage', within which the Limnochromini appeared outside of all other included taxa (Fig. 1g). The study of Friedman et al. (2013), which was based on ten nuclear makers and did not focus specifically on the species of LT but on a larger cichlid phylogeny, revealed a clade formed by the Lamprologini, the Perissodini plus the Cyprichromini, and the Cyphotilapiini plus the Limnochromini as sister group to the Ectodini, the Eretmodini and the Haplochromini (Fig. 1h).

In summary, after more than 20 years of research, the composition of individual LT tribes has been well investigated, whereas the phylogenetic relationships among these cichlid tribes remain largely elusive. All studies performed so far revealed different results (Fig. 1), and the support values for many of the deeper nodes were consistently low. While there is consensus about the position of *T. polylepis*, *O. tanganicae*, the Bathybatini, Boulengerochromini and Trematocarini outside of the other tribes, the following main areas of uncertainty persist: (i) the relative position of the Bathybatini, Boulengerochromini and Trematocarini to each other; (ii) the placement of the Eretmodini, which were suggested as either being part of the 'H-lineage' and sister to the Haplochromini (Friedman et al., 2013; Nishida, 1991; Sturmbauer and Meyer, 1993), as sister-group to the Lamprologini (Clabaut et al., 2005; Day et al., 2008; Kocher et al., 1995), or as separate lineage outside the Lamprologini-'C-lineage' clade (Muschick et al., 2012; Salzburger et al.,

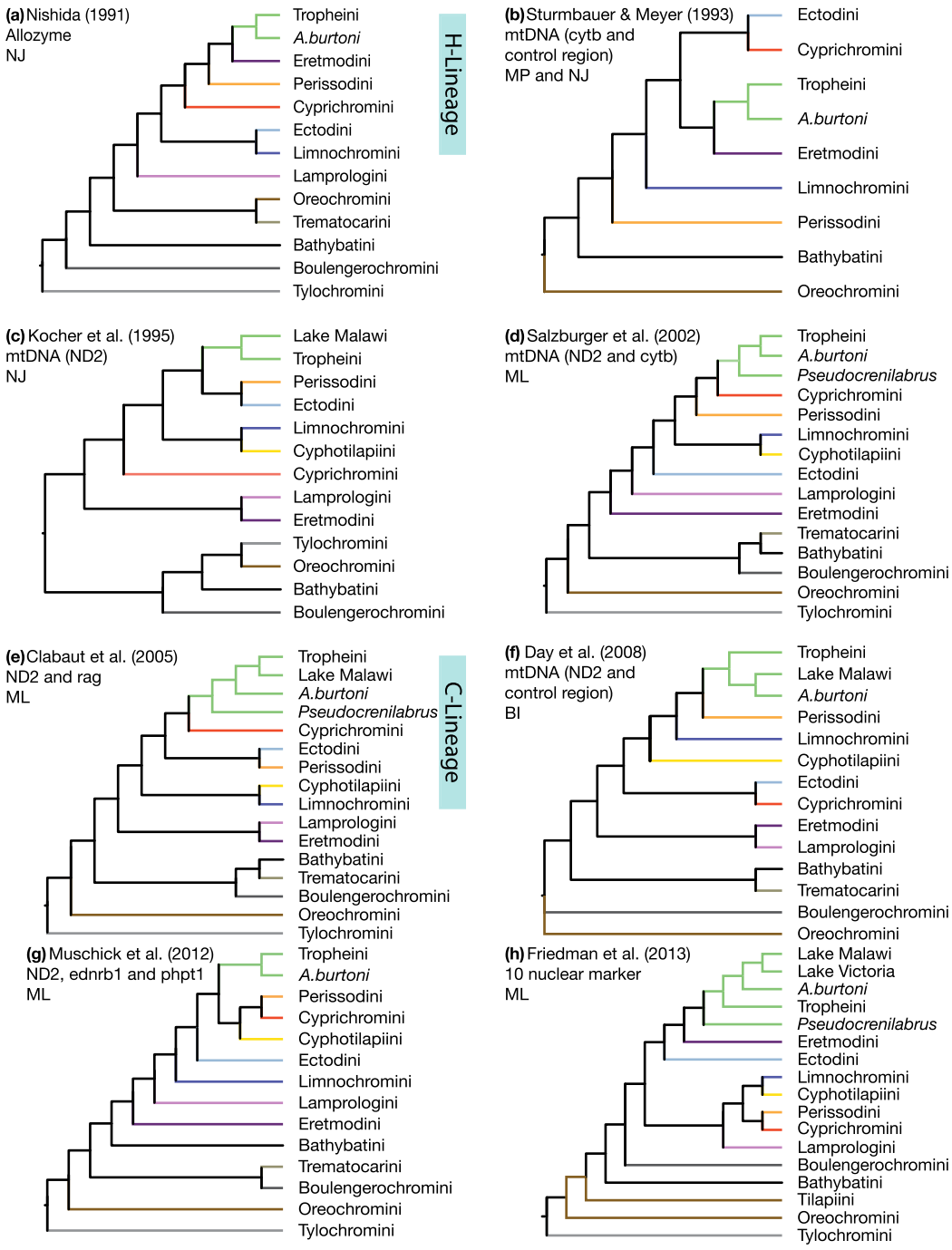

Fig. 1. Previous hypotheses for the phylogenetic relationships among cichlid tribes in Lake Tanganyika. The figure depicts simplified cladograms based on the studies of (a) Nishida (1991), (b) Sturmbauer and Meyer (1993), (c) Kocher et al. (1995), (d) Salzburger et al. (2002a), (e) Clabaut et al. (2005), (f) Day et al. (2008), (g) Muschick et al. (2012), and (h) Friedman et al. (2013). The markers used in the respective study and the phylogenetic algorithms applied are indicated; the color code for cichlid tribes follows that of Muschick et al. (2012). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

2002a); and (iii) the relative position of the 'H-lineage'/'C-lineage' taxa with respect to each other.

The apparent intricacy with resolving the phylogenetic relationships of the cichlid tribes in LT might have various reasons. First, the conflict between the various phylogenetic hypotheses might in part result from the different phylogenetic algorithms used (see above), although this would not apply to the more recent studies, all of which relied on ML and BI methods. Second, we might face the problem here that the previously used markers do not provide enough power of resolution for the question at hand. Alternatively, the inability to resolve some of the phylogenetic relationships of LT's cichlid tribes might reflect biological reality in the context of an adaptive radiation, where speciation is not necessarily bifurcating and multiple lineages may evolve nearly contemporaneously from a common ancestor ('soft polytomy' versus 'hard polytomy' problem: Maddison, 1989; Slowinski, 2001; Sturmbauer et al., 2003; Walsh et al., 1999; Whitfield and Lockhart, 2007). Conflicting topologies may also be the result of reticulate evolution due to (introgressive) hybridization, which is a commonly observed phenomenon in LT's cichlid assemblage (e.g. Koblmüller et al., 2007; Salzburger et al., 2002b) and might have acted as trigger of cichlid adaptive radiations in the first place (Joyce et al., 2011; Seehausen,

2004). Finally, discordance between different sets of markers could reflect incomplete lineage sorting, which is expected to have a strong impact on phylogenetic inference in rapidly diversifying clades (Kubatko and Degnan, 2007) and has been demonstrated in LT cichlid fishes before (Takahashi et al., 2001).

With decreasing sequencing costs and increasing computational resources, single marker and mtDNA-based phylogenies are rapidly being replaced by phylogenies inferred from large-scale nuclear marker sets based on selected loci, transcriptomes, or even whole genomes (McCormack et al., 2013). This recent development enables comparisons between the phylogenetic histories of multiple sets of individual markers. Here, we analyze the phylogenetic history of cichlid fishes from LT on a tribal level, including representatives from the East African Lakes Victoria and Malawi. We sampled 45 species and 42 nuclear loci and thus assembled the largest DNA sequence dataset available for LT cichlid fishes to date. In order to account for potential hybridization and incomplete lineage sorting, we explore gene tree concordance in addition to concatenation as ways for species tree estimation. We further test the strength of our dataset using random resampling of different numbers of markers.

2. Material and methods

2.1. Sample collection and DNA extraction

Specimens for this study were collected between 2007 and 2011 at the Kafue River (Kafue National Park) and at LT in the Northern Province of the Republic of Zambia following the standard operating procedure described in Muschick et al. (2012). Additional samples were obtained from aquaria stocks at the University of Basel and at EAWAG, Kastanienbaum, Switzerland. In total, we analyzed data for 45 specimens, each representing a different East African cichlid species. Our sampling comprised 34 cichlid species from LT covering all major cichlid lineages in this lake. In addition we included 11 further species of riverine clades and from Lakes Victoria and Malawi, to place the LT cichlid taxa into a larger phylogenetic context. A detailed list of specimens, their IDs and sample locations is provided in Table S1. Genomic DNA was extracted from ethanol preserved tissue of whole specimens (see Muschick et al., 2012 for details).

2.2. Marker selection, sequencing and quality control

To infer the phylogenetic history of the cichlid fishes of LT on the basis of an informative set of nuclear (nc) DNA markers, we selected a set of 42 nuclear loci. Twenty-four primer pairs were taken from earlier studies (Meyer and Salzburger, 2012; Muschick et al., 2012; Won et al., 2005) and 18 primer pairs were newly designed following the strategy described in Meyer and Salzburger (2012). In short, we selected genes with known functions and aimed for amplification products between 400 and 600 bp in length to enable the application of next-generation amplicon sequencing. Twenty-four of the markers were developed as exon-primed intron crossing (so-called EPIC) primers (Lessa, 1992; Slade et al., 1993). The markers for *enc1*, *ptr*, *tbr* and *snx33* were taken from Li et al. (2007), but modified to meet our requirements. The same strategy was applied for *ednrb* (Lang et al., 2006), *bmp4* (Albertson et al., 2003), and the reverse primer of *s7* (Chow and Hazama, 1998). The genome of the Nile Tilapia (*Oreochromis niloticus*) (Brawand et al., 2014) was used to define exon–intron boundaries and UTRs. A detailed list of all primers, their base composition, the length of the amplification products, their source, the ENSEBML reference of the respective locus in Tilapia, the chromosomal position of the respective locus in the Medaka genome and the number of variable sites are provided in Tables 1 and 2.

The 42 nuclear markers were PCR amplified in several separate multiplex reactions in a final volume of 25 μL on a Veriti or 2720 thermal cycler (both Applied Biosystems, Rotkreuz, Switzerland). All PCR reactions contained the Multiplex PCR Kit (QIAGEN, Hombrechtikon, Switzerland) and a primer mix including eight to ten barcoded primer pairs (0.1 μM of each primer), water, and template DNA (5–20 ng/μL). We used barcoded fusion primers synthesized by Microsynth (Balgach, Switzerland). The PCR conditions were standardized for all reactions with an initial heat activation phase of 95 °C for 15 min, followed by 35 amplification cycles with denaturation steps at 94 °C for 30 s, annealing steps at 60–62 °C for 90 s and extension steps at 72 °C for 90 s; reactions were completed by a final extension phase at 72 °C for 10 min.

To remove small fragments, residual primers and primer-dimers, we applied the Agencourt AMPure XP magnetic bead system following the manufacturer's protocol (Beckman Coulter, Nyon, Switzerland) and using a bead/DNA ratio of 1:1. Purification results were inspected with a 2100 Bioanalyzer (Agilent, Basel, Switzerland) using the DNA 1000 Kit. The amplification products of five individual PCR reactions with different primer combinations were then pooled (on the basis of the concentration measurements with the Bioanalyzer) to obtain the final libraries containing all 42 markers of one individual. In a second pooling step, 16 barcoded individuals were pooled for one 1/16th run on a 454 PicoTiterPlate. The subsequent library handling and sequencing was conducted by Microsynth (Balgach, Switzerland) with the GS FLX system (454 Sequencing, Roche). Sequencing was unidirectional starting at the forward primer, which also contained the barcodes.

Individual sequences (in both fasta and fastq format) were separated and extracted with Roche's sffinfo tool (described in 454 Sequencing System Software Manual Version 2.6). Quality control was conducted with the software PRINSEQ (v0.20.3) (Schmieder and Edwards, 2011). We excluded individual reads that were shorter than 150 bp, that had an average Phred quality score below 15, or that contained more than 1% unidentified bases coded as "N". In a second step, we filtered out exact duplicates. The assembly to reference sequences from the *A. burtoni* genome (Brawand et al., 2014) was performed with the software bwa and the BWA-SW algorithm (the Burrows-Wheeler Aligner's Smith-Waterman Alignment) (Li and Durbin, 2010). The resulting SAM files were imported into Geneious (v6.1.6–7.0.3, Biomatters Ltd, Auckland, New Zealand; available from http://www.geneious.com), visually inspected, if necessary reassembled, and further trimmed (we allowed a 0.05 error probability limit and a maximum of 10 low quality bases at the 3′ end). The final consensus sequences for each individual and marker were constructed with a 50% threshold, where bases were called "N" if the Phred score was below 20. Sequence data has been deposited on GenBank under the accession numbers KP129679-KP131427 (see Table S2 for details) and KM263618-KM263752 (Santos et al., 2014).

2.3. Alignment and sequence characterization

Sequences for each locus were aligned with the software MAFFT (v7.017) (Katoh and Standley, 2013), using the "—auto" option. Resulting alignments were visually inspected and manually improved when obvious sequencing artefacts (e.g. homopolymers) were observed or homology appeared questionable.

Overall mean distance for each locus was calculated with the software MEGA (v5.2.1) (Tamura et al., 2011) as the total number of differences and the *p*-distance. This was done for all ingroup taxa (i.e. excluding *Tylochromis polylepis*), with pairwise deletion for missing and ambiguous data. For the concatenated alignment the within group mean distance was also calculated for the three most species-rich lineages, the Haplochromini, the Lamprologini and the Ectodini.

Table 1
List of the 42 markers used in this study. The marker name, the forward and reverse sequence of each primer, the Ensembl Gene-ID for the respective locus in Tilapia, the link to the Ensemble entry for Tilapia, the chromosomal position of each locus in Medaka and the reference for the primer sequences are provided.

Name (synonym)	Forward primer [5′-3′]	Reverse primer [5′-3′]	Ensembl-Gene-ID	Link to Ensembl	Chr Medaka	Reference
rag1	TCGGCGCTTTCGGTACGATGTG	TGCCCCTGAAGTGGAASSGA	ENSONIG00000014593	RAG1	6	Meyer and Salzburger (2012)
b2m	GCCACGTGAGTRATTTCCACCCC	ACGCTAYACRGYGGACYCTGA	ENSONIG00000014176	B2M	23	Meyer and Salzburger (2012)
gapdhs	CCCTGGCCAAAGTCATCCACGATA	CACCACTGACACATCGGCCACT	ENSONIG00000007262	GAPDHS	16	Meyer and Salzburger (2012)
Ptchd4	GCGGGTAGTGAATGTGAGTGCG	ACCCAAGACACCCAGCTCCA	ENSONIG00000006708	PTCHD4	24	Meyer and Salzburger (2012)
enc1	CRGTTCGCCTTGCGCTRTTGC	TGGGTGCCGCCTTTGACCAT	ENSONIG00000020511	ENC1	12	Meyer and Salzburger (2012)
phpt1	AGCAGGGTTGACCTTCTCAA	TGGCTAAAATCCCCGATGTA	ENSONIG00000002175	novel gene	4	Muschick et al. (2012)
rps7	CGTGCCATTTTACTCTGGACTKGC	AACTCGTCYGGCTTCTCGCC	ENSONIG00000018698	RPS7	24	Meyer and Salzburger (2012)
tbr1	ATCGTGCCGGGTGCGAGATA	AGGACGGCGTCTCAATCCAGCT	ENSONIG00000008933	TBR1	21	This study
aqp1a.1	ATCAACCCTGCTCGCTCCTTCG	TGCATCGTTGCCTCCGTTGACG	ENSONIG00000009446	novel gene	17	This study
hprt1	TCAGYGATGAGGAGCAGGGTTATG	CGACCGTCATTGGGATGGAGC	ENSONIG00000017584	HPRT1	10	This study
anxa4	TGGACGAGGCCCAGGCTATTCAAG	ACGTCTTCCAGGCAGCCAGACA	ENSONIG00000003465	ANXA4	12	This study
pgk1	CGGTACCTCCCTGTATGACGAGGA	GCAGCCAGATTTGGTCACCTCGA	ENSONIG00000017337	PGK1	14	This study
bmp4	GAGGACCCATGCCCATTCGTTT	GCCACTATCCAGTCATTCCAGCC	ENSONIG00000001366	BMP4	22	Meyer and Salzburger (2012)
bmp2	AGGCCCTGGCCAGCCTAAAA	TCCTGCGTCTGTGGGCATCCTT	ENSONIG00000000958	BMP2	24	Meyer and Salzburger (2012)
TMO-4C4	TTATGCTGAGGTGTTTGGCCTAC	CCACAGCACCCTCCTCATAAAT	ENSONIG00000017439	novel gene	–	This study
fgf6b	CGCAAAGGTGCCACTACAG	TCGCACTGCACGGATGCAAA	ENSONIG00000000017	FGF6 (2 of 2)	23	Meyer and Salzburger (2012)
runx2	CGGGGTTGGTGTTTGAGGGCAA	GCTGACATGGTGTCACTGTGCTGA	ENSONIG00000001025	RUNX2	24	Meyer and Salzburger (2012)
furina	GCTGCATGGGGACAGACAGTCA	ATAGTCACTGGCACCCGCCACA	ENSONIG00000005696	FURIN (1 of 2)	3	Meyer and Salzburger (2012)
wnt7b	GCGTCTCGGGATCCTGTACCACTA	TGCAGGTAAACACCTCCGTCCT	ENSONIG00000008839	WNT7B	6	This study
pax9	TCCCACGGCTGTGTCAGYAA	ACAGAGTGCGAGGAAGGCCA	ENSONIG00000000990	PAX9	–	Meyer and Salzburger (2012)
sox10b	TSCRGGGTCTGGGAAACCTCAT	TGGTGGTCGGCGTATTCTGCAA	ENSONIG00000008392	SOX10 (1 of 2)	8	Meyer and Salzburger, 2012
otx2	GCAGAACAAAGTGCGACCTGCC	GTCTGCTGTGGAGTTGAAGCCCA	ENSONIG00000020156	OTX2	22	This study
otx1	TACACCTCCTGCTGTCTCCAGCAC	ATAGATGAGGCCGTCATGGGGC	ENSONIG00000001278	OTX1 (1 of 2)	15	This study
dlx2a	ATCGCCAACTCCCGCAGACA	TCCGTTGAAGYGCAGCCAGT	ENSONIG00000008722	DLX2	21	This study
dlx4b	GCGTGGATTTCTTCCAGGCTGTC	CTGTGTGCTCTAATCTGCTGTGGG	ENSONIG00000019896	DLX4 (1 of 2)	19	This study
barx1	TCTCGCAGAGTCTCTCGGTCTG	TCGCTGCTGGGGATGGAGTT	ENSONIG00000003234	BARX1	–	This study
ednrb1a	CGTTGGCCTGCACTGCCATT	AGGCAGCCAGCACAGAGCAAA	ENSONIG00000018701	EDNRB (1 of 2)	17	Meyer and Salzburger (2012)
mc1r	GACCACGGCCTCCTGGATGT	GTTGCAGAAGGGGCTGGTGG	ENSONIG00000021393	MC1R	3	Meyer and Salzburger (2012)
skia	CGACCAGCTGGAGATCCT	TCCTCTTGTACTTGTTGGCG	ENSONIG00000017935	SKI (1 of 2)	7	Meyer and Salzburger (2012)
kita	CAGAGTACTGCTGTTTCGGMGAT	GGCTAAGAACTCCATGCCTTTGGC	ENSONIG00000002981	KIT (1 of 2)	4	Meyer and Salzburger (2012)
mitfa	CCTGGCATGAAGCARGTACTGGAC	TTGCYAGAGCACGAACTTCRGC	ENSONIG00000020270	MITF (2 of 2)	5	Meyer and Salzburger (2012)
tyr	TGGGTGGACGCAACTCCCTT	TGGCAAATCGGTCCATGGGT	ENSONIT00000006471	TYR (1 of 2)	13	Meyer and Salzburger (2012)
hagoromo (fbxw4)	AAACTGGTACARYGGGVTCTGC	AGCGRCAGACGTCACCCTTGT	ENSONIG00000013182	HAGOROMO	15	Meyer and Salzburger (2012)
slc45a2 (aim)	GAGCTATGGACTGGGGTCAC	TGGCTGTTTGACACTTGAGG	ENSONIG00000007610	SLC45A2	12	Won et al. (2005)
rh1	TCGCCTTGGCTGCAATCTGG	ACCATGCGGGTGACTTCCCT	ENSONIG00000021142	RH1	7	This study
opn1mw (lws)	ATTGCTGCTCTTTGGTCCCTGACA	AGCCAGAGGGTGGAAGGCAT	ENSONIG00000020292	OPN1MW	5	This study
opn1sw (sws)	TGGGTCACACGCTGTGTGCT	CAGCAGCTGGGAGTAGCAGAARA	ENSONIG00000007620	OPN1SW	scaffold1021	This study
ccng1	CTGCTTGCCCTGGCTCTCCT	AGCTGACTCAGGTATGGTCGGA	ENSONIG00000012912	CCNG1	10	Meyer and Salzburger (2012)
snx33	TGGCTGTACAACCGCCTGCT	CCAAYRTGAATGCSTGGCTGA	ENSONIG00000012857	SNX33	6	This study
rpl13a	ACCTGGCTTTCCTGCGCAAGA	TTGCGAGAGGGCTTCAGACGCA	ENSONIG00000003560	RPL13A	22	This study
edar	TGAGCAGCTGTTGAGCCGCA	CRCATKGCARGYYCTGGCATACA	ENSONIG00000004260	EDAR	21	this study
csf1ra	AAGCACAGATGGGACACGCC	TGTACTGGCCCTGCTCCTGT	ENSONIG00000013065	CSF1R (1 of 2)	10	Meyer and Salzburger (2012)

Table 2
Characterization of the 42 loci used in this study. The marker name, the alignment length of each marker, the sequenced gene regions, the number of variable (*V*) and parsimony informative (PI) sites in the ingroup taxa, the mean number of differences (genetic distance) and the *p*-distance in the ingroup taxa, and the assignment to one of six subsets according to the CONCATERILLAR analysis are specified for each marker.

Name (synonym)	Alignment lengths	Gene regions	V sites ingroup	PI sites ingroup	Genetic distance	*p*-distance	Subset
rag1	418	Exon	49	21	5.10	0.012	1
b2m	478	Exon, intron, UTR	93	50	12.88	0.031	2
gapdhs	458	Exon, intron	57	15	4.35	0.01	4
Ptchd4	394	Exon	32	11	3.59	0.009	4
enc1	376	Exon	21	7	2.95	0.008	5
phpt1	459	Exon, intron	67	31	7.14	0.017	1
rps7	470	UTR	77	31	9.24	0.021	4
tbr1	466	Exon	13	6	1.58	0.003	5
aqp1a.1	440	Exon, intron	62	24	5.69	0.014	2
hprt1	402	Exon, intron	45	14	5.12	0.014	1
anxa4	642	Exon, intron	56	20	6.31	0.014	1
pgk1	377	Exon, intron	40	16	3.55	0.01	3
bmp4	456	Exon	47	16	4.37	0.011	4
bmp2	372	Exon	26	8	1.78	0.005	1
TMO-4C4	428	Intron	54	32	8.02	0.019	2
fgf6b	471	Exon, intron	29	7	2.64	0.006	2
runx2	360	Exon, intron, UTR	16	5	2.06	0.006	1
furina	311	Exon, intron	34	8	2.88	0.009	2
wnt7b	389	Exon	16	4	1.41	0.004	2
pax9	394	Exon	22	7	2.20	0.006	1
sox10b	378	Exon	40	15	4.43	0.012	2
otx2	412	Exon	19	7	1.89	0.005	1
otx1	356	Exon	15	9	1.86	0.005	5
dlx2a	497	Exon, intron	83	27	6.94	0.015	2
dlx4b	356	UTR, exon	29	7	2.43	0.007	4
barx1	220	Exon, intron	30	11	3.47	0.019	1
ednrb1a	438	Exon, intron	59	28	6.82	0.016	6
mc1r	426	Exon	30	9	2.71	0.007	1
skia	453	Exon	38	11	2.67	0.006	2
kita	431	Exon, intron	45	20	4.93	0.012	2
mitfa	434	Exon, intron	57	21	6.41	0.016	6
tyr	525	Exon, intron	72	26	8.47	0.019	3
hagoromo (fbxw4)	493	Exon, intron	110	59	16.01	0.043	2
slc45a2 (aim)	286	Exon	38	16	4.55	0.016	3
rh1	404	Exon	43	32	9.59	0.024	6
opn1mw (lws)	420	Exon, intron	53	22	6.65	0.017	1
opn1sw (sws)	450	Exon, intron	80	36	10.01	0.024	1
ccng1	460	Exon, intron	69	20	6.55	0.017	1
snx33	437	Exon	43	19	5.10	0.012	1
rpl13a	370	Exon, intron	28	9	3.00	0.013	4
edar	372	Exon, intron	41	13	3.29	0.009	2
csf1ra	366	Exon, intron	54	19	5.29	0.015	2

2.4. Gene tree discordance tests

We first tested for topological incongruence between individual gene trees, using hierarchical likelihood ratio tests as implemented in the software CONCATERPILLAR (v1.7.2) (Leigh et al., 2008), with default settings and the assumption of linked branch lengths. As part of the CONCATERPILLAR analysis, tree inference was performed using RAxML (v7.2.8) (Stamatakis, 2006), assuming a single GTR substitution model for each sequence alignment. The two largest sets of markers identified by CONCATERPILLAR to have concordant histories (containing 13 and 14 markers, respectively) were each concatenated and subjected to phylogenetic analyses as described below.

2.5. Phylogenetic analysis of concatenated datasets

In brief, sequence alignments for sets of loci were concatenated according to different strategies (see below) and phylogenetic analyses were based on both maximum likelihood with GARLI-PART (v2.0.1019) (Zwickl, 2006) and RAxML (v7.7) (Stamatakis, 2006), and on Bayesian inference with MrBayes v3.2.1 (Ronquist et al., 2012). Prior to tree inference, sequence alignments were subdivided according to gene region (exons, introns and UTRs) and codon position, and the optimal substitution models and partitioning schemes for these subdivisions were selected with the greedy algorithm of PartitionFinder (v1.1.1) (Lanfear et al., 2012) applying the Bayesian information criterion (BIC), and always taking into account substitution models available in the respective tree inference software (Schwarz, 1978). Phylogenetic analyses were run locally or at the CIPRES Science Gateway (Miller et al., 2010) and at Bioportal (Kumar et al., 2009).

We first inferred the phylogeny for each of the two largest sets of loci with concordant histories according to CONCATERPILLAR. To this end, sequence alignments of all markers included in each set were concatenated. We then used concatenation of the full set of 42 loci to infer the phylogenetic history of LT cichlid fishes. This method assumes that all markers share a common evolutionary history and that discordant signals resulting from homoplasies can be counterbalanced by extensive and genome wide marker sampling (Rokas et al., 2003). While the assumption of a common evolutionary history seems to be violated at least for the analysis of the full marker set, concatenation may still lead to correct phylogenetic estimates when the true tree lies outside of the "anomaly zone" (Kubatko and Degnan, 2007). As there is no fully unlinked branch length option in GARLI, analyses were run with linked branch lengths (subsetspecificrates = 1, linkmodels = 0) and partitioning schemes and substitution models selected by PartitionFinder with respective settings (branchlengths = linked,

models = all, resulting in 17 distinct partitions for the full-concatenated dataset). A total of 50 independent ML inferences were conducted in GARLI, with the termination condition set to at least 10,000 generations without any substantial (0.01) topological enhancement. Node support was assessed with 500 replicates of non-parametric bootstrapping with the same settings. Bootstrap values were mapped to the ML topology with SumTrees (v3.3.1), using the DendroPy Phylogenetic Computing Library (v3.12.0) (Sukumaran and Holder, 2010).

ML phylogenies with unlinked partition-specific branch lengths were estimated with RAxML, using the -M option and applying a partitioning scheme obtained by a PartitionFinder analysis (settings: branchlengths = unlinked, model = raxml, resulting in 2 partitions). For the ML inference, we used RAxML's rapid hill-climbing algorithm and the GTR + GAMMA model in 50 alternative runs and with 500 bootstrap replicates each.

Likewise, MrBayes analyses were conducted with unlinked branch lengths (unlink brlens = (all), prset ratepr = fixed) and a partitioning scheme estimated by PartitionFinder (settings: branchlengths = unlinked, model = mrbayes, resulting in 2 partitions). Using the default prior probability distribution (exponential prior with a mean of 0.1) on branch lengths, two independent MrBayes runs were conducted with four chains for 10,000,000 MCMC generations, sampling every 100th generation, and discarding the first 25% as burn-in. All other settings were left at their defaults. Convergence of MCMC was assessed by MrBayes' Potential Scale Reduction Factor (PSRF) reaching 1.0, and the average standard deviation of split frequencies falling below 0.01. We further evaluated effective sample sizes in Tracer (v1.5) (Rambaut and Drummond, 2007) and plotted posterior probabilities of splits over the MCMC run with AWTY online to test for convergence of runs (Nylander et al., 2008).

To examine the phylogenetic signal contained in length-mutational events and to evaluate the potential power of a combined analysis (alignment plus indel information), the indels from the concatenated alignment were translated into a presence/absence matrix. This was performed with the software SeqState v1.4.1 (Müller, 2005) using the simple indel coding procedure (SIC) (Simmons and Ochoterena, 2000). Phylogenetic inference for these two datasets was conducted with GARLI, applying the Mkv model of Lewis (2001), and otherwise using default settings as described above.

2.6. Gene tree summary statistics and Bayesian concordance analysis

In order to visualize potentially conflicting signal contained in the 42 loci, gene trees for each individual marker were inferred using GARLI with settings as specified in Section 2.5. The 50 best topologies from each run and from all 42 markers (a total of 2100 gene trees) were used to generate an average consensus tree in SplitsTree (v4.12.3) (Huson and Bryant, 2006). The implemented "average consensus tree" function constructs a neighbor-net using the average pairwise distances of the individual trees.

As a further approach to investigate the discordance among the sampled gene trees and to combine conflicting data in a primary concordance and a population tree, we applied a Bayesian concordance analysis (BCA) (Ane et al., 2007; Baum, 2007), as implemented in the software BUCKy v1.4.0 (Larget et al., 2010). Using samples of MrBayes' posterior tree distribution as input, this analysis accounts for both uncertainty in individual gene trees and potential discordance among trees inferred from different loci. The primary concordance tree, as estimated by BUCKy, visualizes the most dominant history from several gene trees, along with concordance factors (CF) indicating the proportion of loci supporting a given clade (Baum, 2007). In addition, a population tree with coalescent units as branch lengths is generated by BUCKy, based on quartets of concordance factors. This population tree is known to be consistent in the presence of incomplete lineage sorting (Chung and Ané, 2011; Larget et al., 2010).

In order to apply BUCKy, MrBayes was used to infer gene trees from the individual loci, with substitution models and partitioning schemes selected by PartitionFinder (assuming linked branch lengths for all subdivisions of each locus). For each locus, we conducted two replicate MrBayes runs with six chains of 15 million generations, sampling every 100th generation. As reported by Willis et al. (2013), we found that for most loci, all of the 150,000 sampled trees represented unique topologies, suggesting a lack of resolution in some parts of the tree. This could partly be due to polytomies, which would be displayed as multiple weakly supported topologies with very short branches in MrBayes, as this software only provides fully resolved trees. To reduce the large number of distinct tree topologies, we pruned our dataset to 14 taxa, keeping only one representative per tribe (as our primary interest was a tribal level phylogeny). This deletion was done with the pruning option in BUCKy. The BUCKy analysis was conducted with 4 runs, 10 chains and 500,000 generations per chain. The alpha prior, which represents the a priori expected level of discordance, was set to 1–100.

2.7. Testing the strength of the phylogenetic signal as a function of dataset size

In order to test whether our dataset contains a sufficiently large number of markers to recover the "true" phylogenetic history of LT cichlids, we randomly resampled and concatenated different numbers of markers, and produced ML phylogenies from these sets. We then measured the topological difference between the tree resulting from one set of randomly chosen markers and the tree resulting from the complete set including of all markers and between the trees resulting from two different and mutually exclusive marker sets. As our full dataset contained 42 markers, the first comparisons were done for 1–41 randomly chosen markers, whereas the latter was performed for 1–21 randomly chosen markers. For each number of markers between 1 and 41, we compiled 20 sets drawn at random from the full set of 42 markers. Then, for each of the sets containing at most 21 markers, a comparison set was produced containing the same number of markers so that the two sets did not share any marker. In order to take into account marker concordance according to the results of the CONCATERPILLAR analysis (see Section 2.4.) we repeated the same procedure for 1–13 markers, again with 20 replications each. For the latter analysis, we always compiled two sets of markers, so that markers shared a concordant history within each set, but a discordant history between the two sets (according to CONCATERPILLAR). All generated marker sets were subjected to phylogenetic analysis with GARLI (see above, Section 2.5.), using marker-specific partitions and substitution models as suggested by PartitionFinder. Topological differences between resulting ML trees were measured by means of their *K*-score (Soria-Carrasco et al., 2007), as the *K*-score accounts for variable substitution rates between marker sets.

Then, *K*-scores of 20 replicate comparisons were plotted against the number of markers used in the datasets for which the respective ML trees had been inferred (see Camargo et al., 2012; Willis et al., 2013). We expected a general decrease of mean *K*-scores (i.e., fewer topological differences) with increasing marker number due to an increase in the phylogenetic signal for larger datasets. We further expected *K*-scores between a tree based on randomly drawn markers and the tree based on the full dataset of 42 markers to approach zero for marker numbers close to 42, as the alignments used for the reconstruction of the two trees would become increasingly similar. Nevertheless, we expected the degree to which *K*-scores decrease with increasing number of markers to inform

about the minimum number of markers needed to reliably construct the relationships among cichlid tribes in LT.

As an additional measure of discordance, we tested for statistically significant topological differences between the tree based on all 42 markers, and trees based on smaller datasets, using the Shimodaira–Hasegawa (Shimodaira and Hasegawa, 1999) and Approximated Unbiased (Shimodaira, 2002) tests as implemented in PAUP* (v.4.0a129) (Swofford, 2003). For each number of markers between 1 and 41, we plotted the number of tree replicates that fitted the full dataset significantly worse than the tree produced from all 42 markers.

3. Results

3.1. Sequencing

Amplicon sequencing was successful for most of the 42 markers for the 45 taxa. In total, we obtained 98.3% of the 1890 possible sequences. Of 789,525 bp in the final alignment, 26,854 bp (3.40%) consisted of gaps; 27,211 bp (3.45%) were undetermined ("N") and 476 bp (0.06%) were ambiguous ("WRYSMK" coded).

3.2. Alignment and sequence characterization

The concatenated alignment had a total length of 17,545 bp, of which 1932 positions (11.01%) were variable and 769 positions (4.38%) were parsimony informative (not considering the outgroup *Tylochromis polylepis*). The amount of variable sites per marker varied between 13 and 110 sites (average: 46, median: 43), the number of parsimony informative sites ranged between four and 59 (average: 18.3, median: 16) (Table 2). The average sequence length for each marker was 417.7 bp (median: 423 bp), and the average total number of differences across all sequence pairs was 208.8 (uncorrected *p*-distance: 0.013). Within three of the major lineages, we found that the Ectodini showed the highest divergence (114.1 differences; uncorrected *p*-distance: 0.007), followed by the Lamprologini (110.4; 0.007) and the Haplochromini (all species included; 103.1; 0.006). Separate analyses of the within group mean distance of the haplochromines of the three lakes indicated a higher number of base differences between the four species of Lake Malawi (14.3; 0.0009) than the four species of Lake Victoria (6.8; 0.0004). The Tropheini (*Ctenochromis horei, Lobochilotes labiatus, Gnathochromis pfefferi, Tropheus moori*) included in this study showed a higher level of diversity (73.0; 0.004).

3.3. Gene tree discordance tests

We used CONCATERPILLAR to test for topological incongruence between markers and to identify concordant sets of markers. Based on hierarchical likelihood ratio tests, CONCATERPILLAR detected six sets of markers that were concordant internally, but exhibited significant levels of discordance (p-value < 0.001) between them. The three largest sets contained 14, 13, and 6 markers, respectively, whereas the remaining three sets included 3 markers each (the assignment of each marker to one of these subsets is indicated in Table 2). The six sets exhibited no obvious clustering of markers according to gene function, coding and non-coding parts, or variability. The two largest sets of markers were subjected to individual phylogenetic analysis. Subset 1 (14 markers) contained a total of 5872 concatenated bp, of which 10.30% were variable and 3.92% were parsimony informative. The average pairwise distance was 61.05 mutational steps, and the uncorrected *p*-distance was 0.012. Subset 2 (13 markers) had a length of 5507 bp, with 12.69% variable sites, and 5.25% parsimony informative sites. This marker-set showed a somewhat higher variability (average pairwise distance: 76.07; uncorrected *p*-distance: 0.015).

3.4. Phylogenetic analysis of concatenated datasets

Phylogenetic analysis of concatenated subsets revealed conflicting topologies between subset 1 and subset 2 (Fig. 2a and b). While the base of the resultant trees (i.e. the position of the Bathybatini, Boulengerochromini and Trematocarini) was highly similar, the topologies differed with respect to the relative placement of the Eretmodini, the Lamprologini, the Limnochromini and the Cyprichromini/Perissodini clade. For subset 1 the three inferred topologies from the different analyses were congruent. In these trees, the Lamprologini were nested within the mouthbrooding tribes of the 'H-lineage', of which the Cyprichromini/Perissodini clade branched off first. The Lamprologini were resolved as sister group to the Limnochromini in BI (BPP 0.81), and the same relationship was weakly supported in GARLI and RAxML inferences (BS 37 and 45). The Ectodini were placed as sister group to a clade formed by Cyphotilapiini, Eretmodini and the Haplochromini (GARLI BS 26, RAxML BS 35, BPP 0.89).

The phylogenetic analyses of subset 2 revealed a monophyletic group containing the 'H-lineage' taxa (BS 99, BPP 1.0), which were placed as sister taxon to the Lamprologini. The Eretmodini branched off first, and the Ectodini were consistently grouped together with a clade formed by Cyprichromini/Perissodini, the Cyphotilapiini and the Limnochromini (BS 32-34, BPP 0.90). In general, the interrelationships of tribes received only moderate support, which is likely a consequence of the comparatively small number of markers in this subset (see Section 3.6). Excluding *Eretmodus cyanostictus* from these two phylogenetic analyses did not change the resulting tree topologies (data not shown).

The trees obtained with the entire concatenated dataset of 42 markers were highly congruent and most nodes were very well supported (mean GARLI BS 79.2; mean RAxML BS 78.1; mean BPP 0.941). Fig. 3b depicts the ML tree inferred with GARLI; the ML tree obtained with RAxML and the 50% majority rule consensus tree of our MrBayes analysis are shown in Fig. S1. In all three trees, *Oreochromis tanganicae* appeared as the sister to *Tilapia sparrmanii* and a strongly supported clade formed by the remaining tribes (GARLI BS 100, RAxML BS, 100, BPP 1.0). The monophyly of these tribes was strongly supported (BS 100, BPP 1.0 for all tribes of which more than two representatives have been included). Within this group *T. nigrifrons* and *B. graueri* appeared as sister taxa (BS 100, BPP 1.0) in all our analyses. The three tribes Boulengerochromini (represented by their only member, *B. microlepis*), Trematocarini (represented by *T. nigrifrons*), and Bathybatini (represented by *B. graueri*) appeared outside of a strongly supported clade (BS 100, BPP 1.0), in which the substrate spawning Lamprologini, the most species-rich tribe within LT, are clearly separated from the mouthbrooding tribes (i.e. Cyphotilapiini, Cyprichromini, Ectodini, Eretmodini, Haplochromini, Limnochromini, Perissodini; BS 73-75, BPP 1.0).

The branching order within the mouthbrooding tribes of the 'H-lineage' received less support, and there was incongruence between the tree topologies resulting from the different analyses with respect to the placement of the Cyphotilapiini and the Limnochromini relative to each other, and regarding the first divergence events within the Haplochromini (indicated by dotted lines in Fig. 3b). The Cyprichromini were consistently resolved as the sister group of Perissodini (BS 100, BPP 1.0), and the clade formed by these two tribes represented the sister of all remaining tribes of the 'H-lineage' in all analyses of the full-concatenated dataset. The Limnochromini and the Cyphotilapiini formed a monophyletic group that was sister to a clade combining the Ectodini, the Eretmodini, and the Haplochromini (GARLI BS 65, RAxML BS 59, BPP

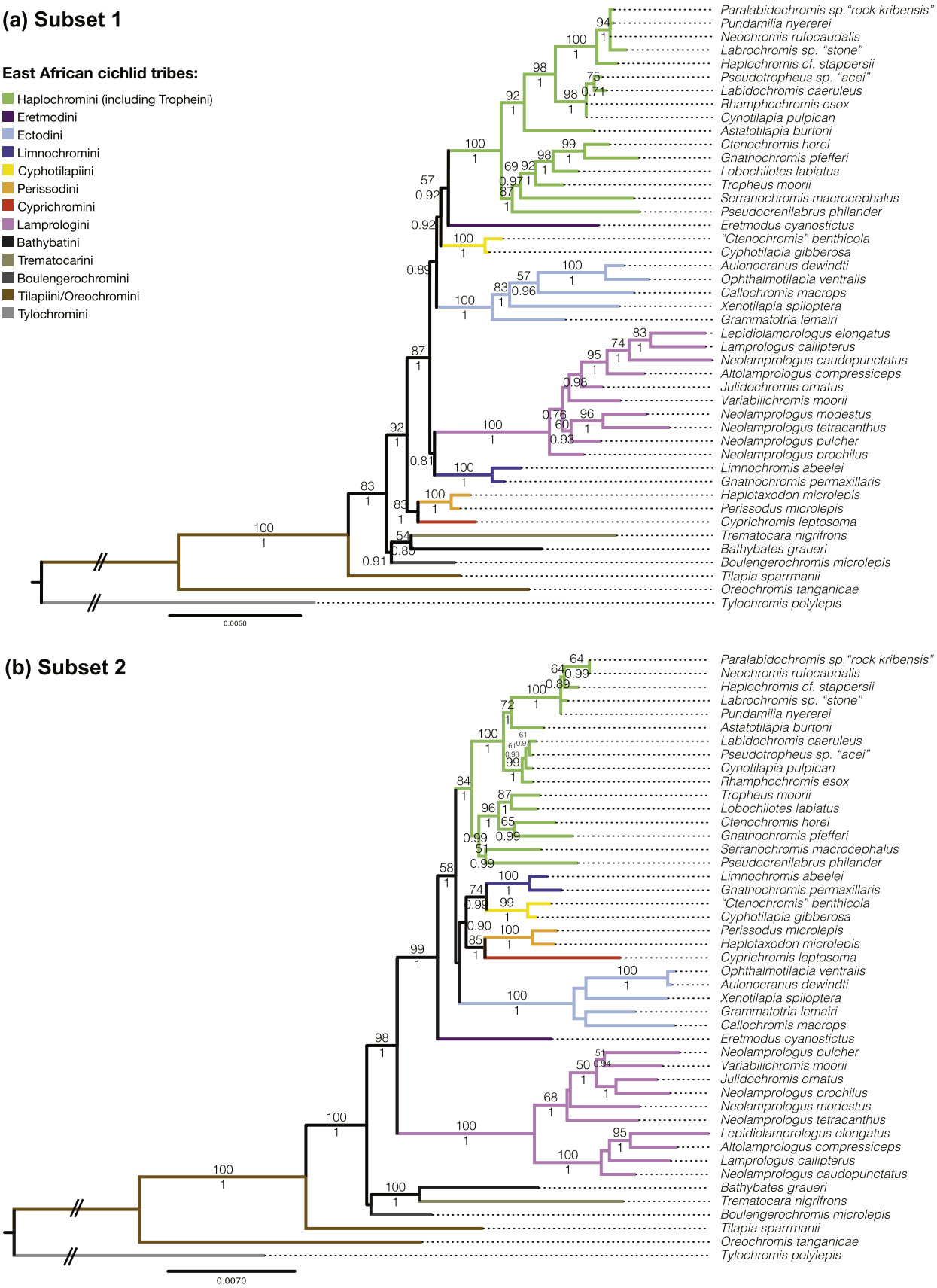

Fig. 2. Results from the phylogenetic analyses based on the two largest subsets of markers identified with CONCATERPILLAR. (a) Maximum likelihood phylogeny of subset 1 (14 markers; see Table 2) inferred with GARLI. (b) Maximum likelihood phylogeny of subset 2 (13 markers; see Table 2) inferred with GARLI. Numbers above the branches represent maximum likelihood bootstrap support values (⩾50%) as obtained with GARLI, numbers below the branches indicate Bayesian posterior probabilities (⩾0.75) as revealed with MrBayes. The branch leading to the outgroup taxon, *Tylochromis polylepis*, is shortened by one third. The colors indicate the affiliation of each taxon to one of the cichlid tribes.

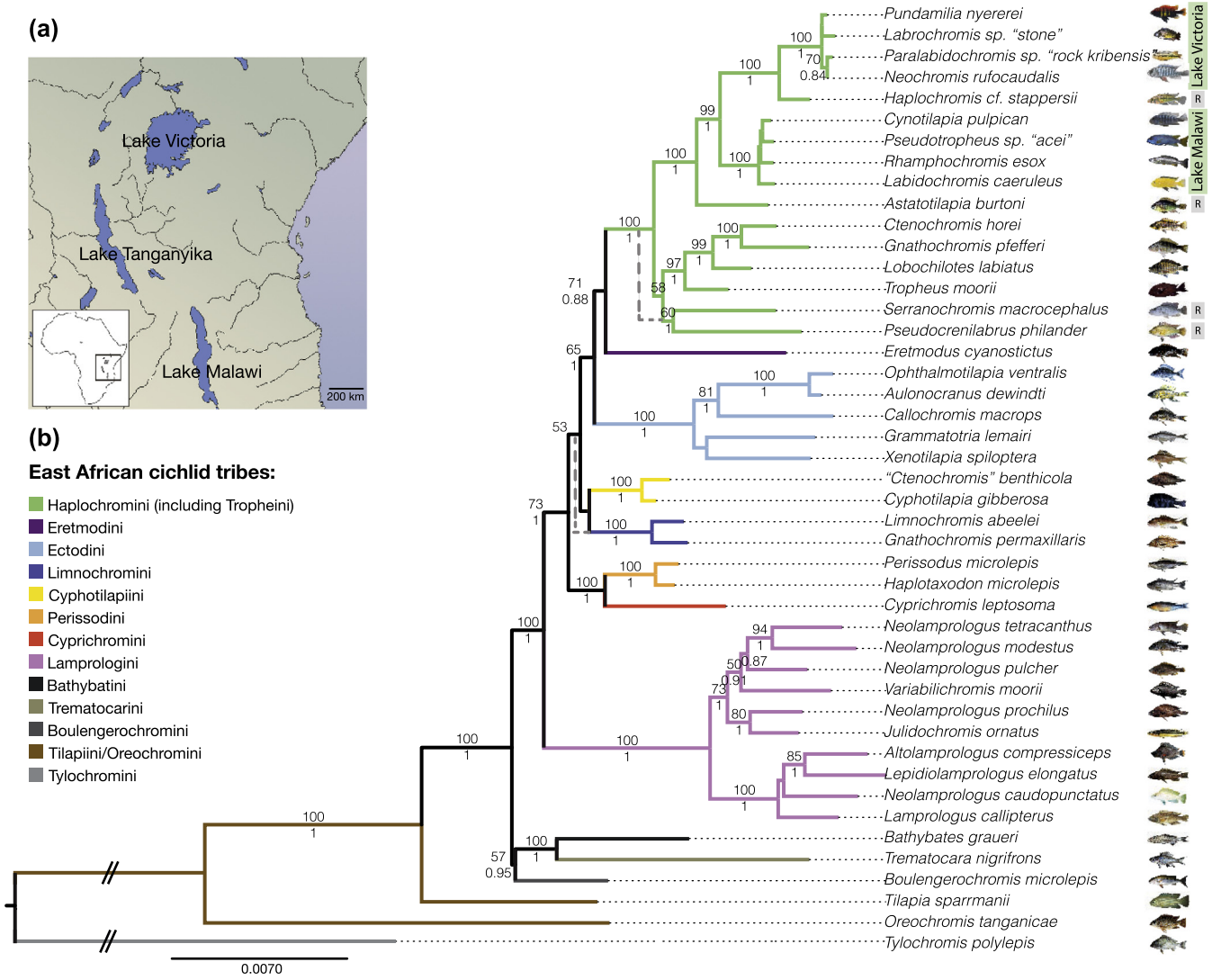

Fig. 3. Tribal level phylogeny of the Lake Tanganyika cichlid fishes. (a) Map of the area showing the three East African Great Lakes. (b) Maximum likelihood tree based on the concatenated dataset (17,545 bp) as obtained from a partitioned analysis with GARLI. Numbers above the branches indicate maximum likelihood bootstrap support values (⩾50%) produced with GARLI, numbers below the branches represent Bayesian posterior probabilities (⩾0.75) as revealed with MrBayes. Alternative branching orders between the maximum likelihood analysis with GARLI (as shown here) and the maximum likelihood analysis with RAxML (Fig. S1a) and Bayesian inference with MrBayes (Fig. S1b) are indicated with dotted lines; the branch leading to *Tylochromis polylepis* was shortened by one third; colors indicate the tribal affiliation of each taxon. Sample origin other than LT are indicated with boxes on the right; R = riverine. Fish pictures were taken in the field, except for *P. nyererei* and *R. esox* (credit: E. Schraml), *P. rockkribensis* (credit: M. Negrini) and *L. sp.* 'stone' (credit: O. Seehausen).

1.00) in the GARLI analysis, whereas the Cyphotilapiini appeared closer to this clade according to the RAxML and MrBayes analyses. Within this clade, the representative of the Eretmodini (*E. cyanostictus*) was consistently placed as sister group to the Haplochromini (GARLI BS 71, RAxML BS 50, BPP 0.88). Similarly, the species from Lake Victoria and Lake Malawi appeared reciprocally monophyletic (BS 100, BPP 1.0) within the Haplochromini. *Haplochromis cf. stappersii* from LT was resolved as sister taxon to the Lake Victoria cichlids (BS 100, BPP 1.0). The riverine species *Astatotilapia burtoni* was always placed outside of the species-flocks of the Lake Malawi and Victoria cichlids (BS 100, BPP 1.0). The haplochromines *Serranochromis macrocephalus* and *Pseudocrenilabrus philander* were either put into a separate clade (in RAxML and BI), or placed together with the LT haplochromines (Tropheini) (with GARLI).

Translating all indels of the 42 loci into a binary code resulted in a dataset comprising 167 positions, of which 70 were parsimony informative. A phylogenetic hypothesis obtained with this dataset with GARLI was, overall, concordant with the trees resulting from the concatenated dataset. However, while the monophyly of most tribes and the position of the Eretmodini as sister group to the Haplochromini was recovered, the respective support values were generally low and the position of most of the tribes relative to each other could not be recovered (see Fig. S2).

3.5. Gene tree summary statistics and Bayesian concordance analysis

Inferring single gene trees from 42 genes and 45 taxa with both GARLI and MrBayes (data not shown) resulted in 42 alternative topologies with some to numerous polytomies or low support values for certain branches, whereas other parts of the trees were well resolved. Fig. 4 shows the average consensus network of 2100 trees with 168 splits representing the conflicting affinities within the individual gene trees at the base of the tribes. The tribes themselves seem clearly defined and show only few alternative splits.

For the Bayesian concordance analysis with BUCKy, we pruned the dataset to one representative per tribe (Fig. 5). Changes in the alpha prior had no influence in the topology of both primary concordance and population tree. Its topology (with alpha default

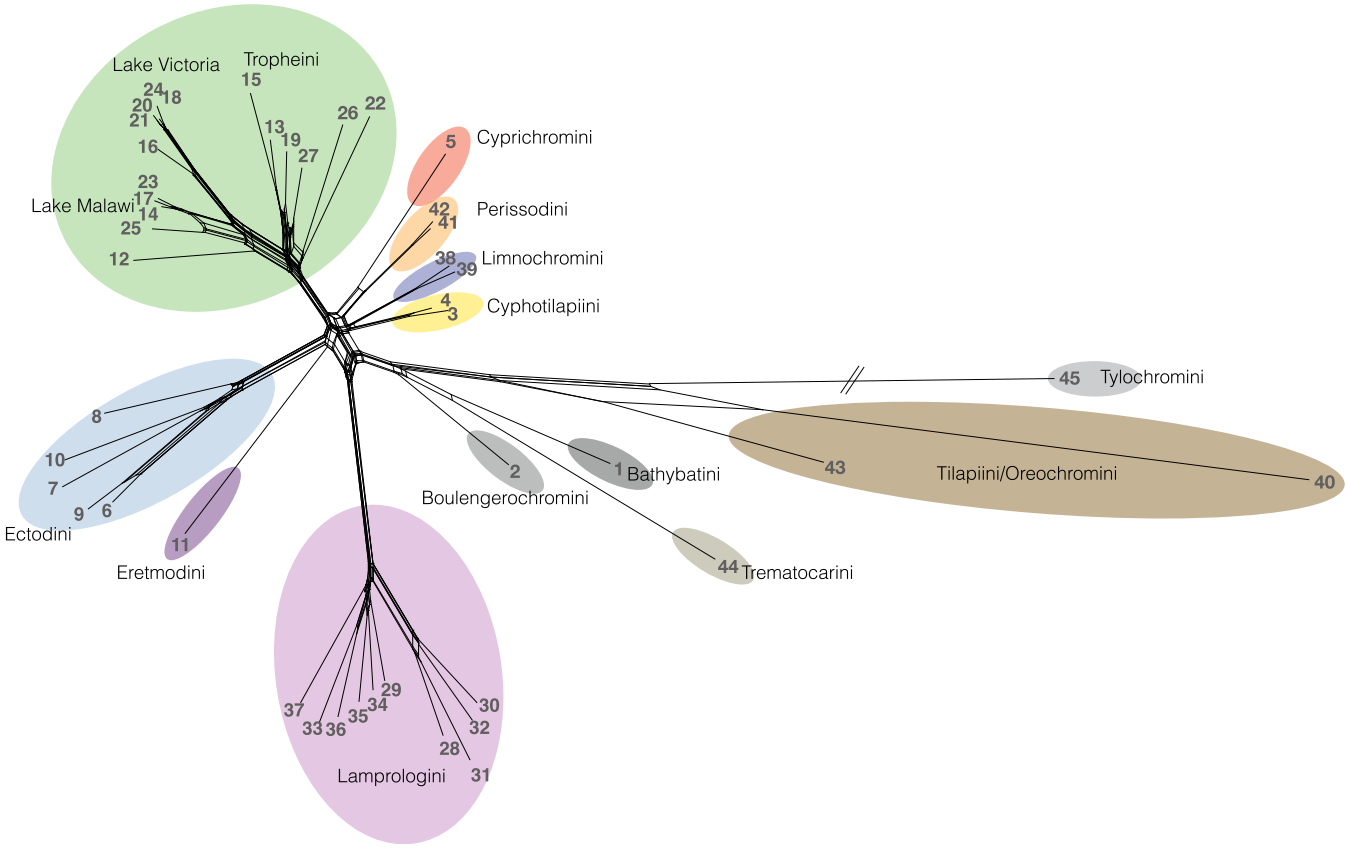

Fig. 4. Average consensus neighbor-net inferred with SplitsTree4 from average pairwise distances in the best gene trees obtained from 50 GARLI runs for each marker (2,100 trees). Note that in this consensus network each gene tree estimate contributed equally and that differences in alignment lengths, degrees of variation, and uncertainties (e.g. bootstrap values) among markers are not considered. The color code is the same is all other figures, the numbers refer to the different species (see Table S1). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

prior) is mostly consistent with the species tree inferred from the full-concatenated dataset (see above; Fig. 3). However, one topological disagreement was found regarding the position of *Boulengerochromis microlepis*, which was placed as a sister group to the clade composed of the Lamprologini and the representatives of the 'H-lineage' (including the Eretmodini) in the population tree, but clustered with the Trematocarini and the Bathybatini in the primary concordance tree. Within the population tree the Eretmodini were again resolved as sister group to the Haplochromini. This close relationship is also reflected in the concordance factors of splits within the primary concordance tree (see Text S1).

3.6. Strength of the phylogenetic signal as a function of dataset size

After 20 repetitions of random resampling and concatenation of 1–41 markers, we used GARLI to infer ML phylogenies from all replicate marker sets, and compared the resulting trees between each other and with the optimal tree based on the full concatenated dataset of 42 markers, in order to test the strength of the phylogenetic signal as a function of dataset size. We expected topological differences between two trees to decrease with increasing size of the respective marker sets as shown in Camargo et al. (2012). Different types of comparisons were performed: Between one tree based on 1–41 markers and the tree resulting from the full marker set (Fig. 6a), between two trees produced from mutually exclusive sets containing 1–21 markers (Fig. 6b), and between two trees based on mutually exclusive sets of 1–13 markers found to be internally concordant but externally discordant in topology according to the CONCATERPILLAR analysis (Fig. 6c).

As expected, topological differences between two trees, as measured by their *K*-score, generally decreased with increasing marker number; the steepest decrease was observed for marker numbers between 1 and 8–10. The median *K*-score between one tree based on a randomly compiled marker set of a given size and the tree based on the full set of 42 markers was always lower than median *K*-scores between two trees based on randomly compiled marker sets of the same size (Fig. 6a versus b). Furthermore, topological comparisons involving the tree based on the full marker set generally resulted in a lower variance of *K*-scores than comparisons between two trees that were produced from randomly sampled mutually exclusive marker sets. In the latter case, the two trees represent independent phylogenetic estimates and are thus particularly useful to assess variance in discordance as a function of marker set size. For this type of comparisons, *K*-scores appear relatively constant for datasets combining between 11 and 21 markers. Nevertheless, *K*-scores between trees based on 21 markers (mean 0.0111) are significantly lower than those between trees constructed from sets of 16 markers (mean 0.0140, *t*-test *p*-value = 0.01613) or less (means $\geqslant$ 0.0128, *t*-test *p*-values $\leqslant$ 0.01704). For most marker set sizes, mean and median *K*-scores of two trees based on mutually exclusive marker sets were slightly lower when all markers with a set were concordant according to the CONCATERPILLAR analysis (Fig. 6c) compared to when sets were composed of randomly sampled markers (Fig. 6b). This reduction was significant for marker sets with eight markers or more (*t*-test *p*-values $\leqslant$ 0.0295), with the exception of sets

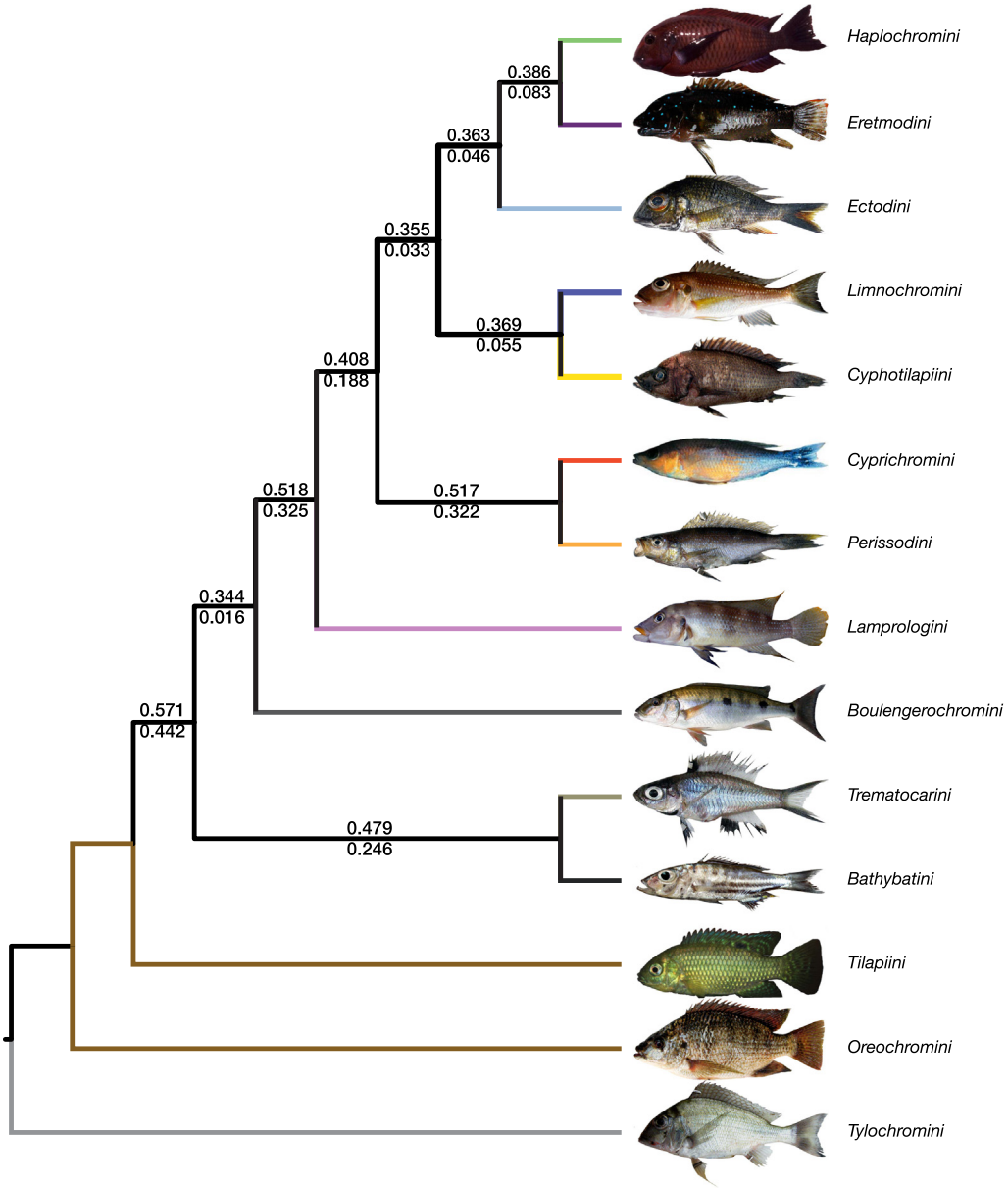

Fig. 5. Population tree topology from the Bayesian concordance analysis (conducted with BUCKy) of 14 taxa representing the different cichlid tribes in LT. Numbers above the branches represent the averaged concordance factors, numbers below are coalescence units (see Text S1 for further details). Fish pictures and color codes are the same as in Fig. 4. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

containing eleven markers (*t*-test *p*-value = 0.0881), which suggests that the discordance between the two largest marker sets identified by CONCATERPILLAR is lower than that between randomly compiled marker sets of the same size.

Similarly, the number of marker set replicates, for which ML trees differ significantly from the ML tree based on 42 markers, shows an overall decrease with increasing size of the respective marker sets. For concatenated sets of 1–5 markers, and for sets of 8 markers, phylogenies produced from all 20 replicate sets are significantly different to the full ML tree, according to both the SH and the AU tests. On the other hand, for concatenated sets of 34 or more markers, none of the phylogenies based on these sets differ significantly from the tree obtained with the full set of markers, according to either of the two tests. Between these extremes, we observe a general decrease in the number of rejected tree replicates with increasing number of markers, based on which these trees were produced (Fig. 6a).

4. Discussion

The present study is the most extensive phylogenetic analysis of cichlid fishes in East African Lake Tanganyika with respect to the number of nuclear DNA markers and the total length of the ncDNA sequences analyzed. The main goal of our work was to establish a robust phylogenetic hypothesis for the relationships among the cichlid tribes of LT, which has so far been inferred on the basis of mtDNA or relatively few nuclear markers only (Clabaut et al., 2005; Day et al., 2008; Friedman et al., 2013; Kocher et al., 1995; Muschick et al., 2012; Salzburger et al., 2002a; Sturmbauer and Meyer, 1993).

The comparatively high information content provided by mtDNA sequences and the availability of universal primers were the main reasons for the utilization of mtDNA markers in earlier phylogenetic analyses aiming to resolve the relatively young and rapid radiation of cichlid fishes in LT. Among the many drawbacks

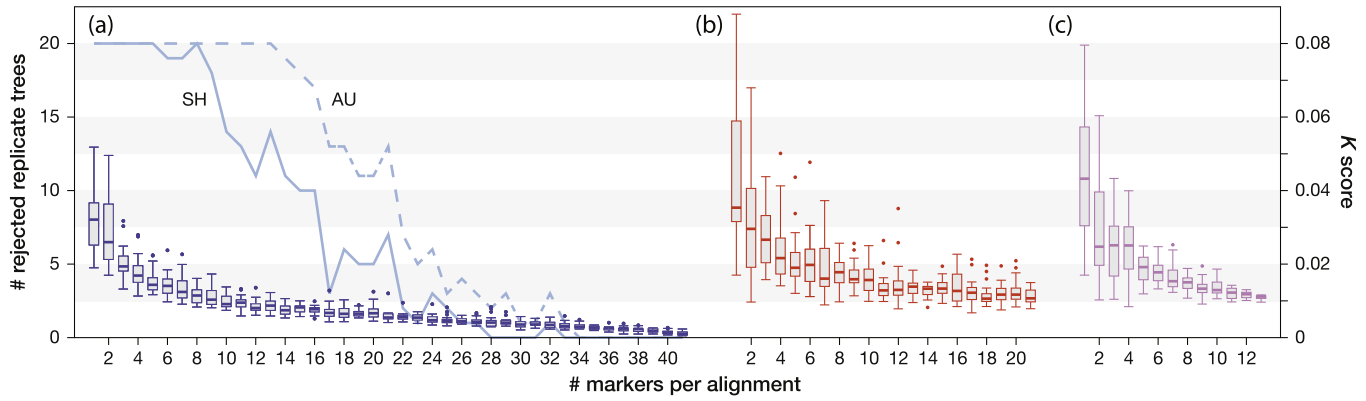

Fig. 6. Topological differences between ML trees measured by their *K*-scores as a function of the number of randomly resampled and concatenated markers. (a) *K*-scores between trees based on randomly sampled and concatenated markers and the tree based on the full dataset of 42 markers. Light blue lines indicate the number of tree replicates (out of a total of 20 replicates) significantly different to the tree based on the full dataset, according to the Shimodaira–Hasegawa (SH) test (solid line), and the Approximately Unbiased (AU) test (dashed line). (b) *K*-scores between two trees that are both based on mutually exclusive randomly sampled marker sets of the given size. (c) As (b), but strictly grouping concordant markers in each set (according to CONCATERPILLAR, see text). Boxplots are based on 20 replicates of each comparison. Whiskers indicate the lowest *K*-score still within 1.5 inter-quartile range of the lower quartile, and the highest *K*-score still within 1.5 inter-quartile range of the upper quartile. Outliers are indicated with dots. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

of mtDNA markers are that only maternal inheritance patterns are captured and that past events of introgression and hybridization remain largely invisible (Ballard and Whitlock, 2004). In addition, a single locus (irrespective of being based on mtDNA or ncDNA) might not accurately reflect the species tree, as individual gene trees often differ from the true species tree (Pamilo and Nei, 1988). Nuclear DNA markers, on the other hand, usually contain fewer variable sites thus less phylogenetic signal. Clabaut et al. (2005) showed, for example, that in LT cichlids, ncDNA datasets would need to contain about ten times more sequence data to obtain the same quantity of phylogenetic information as provided by mtDNA markers – a task not reached by any previous study.

Here we took advantage of the 454 next-generation pyrosequencing technology and compiled a ncDNA dataset for LT cichlids containing 42 markers in well characterized genes and reaching a total alignment length of 17,545 bp. We chose a locus re-sequencing strategy with barcoded primers in order to obtain long enough sequence reads and to sample a large number of gene histories. Primers were chosen to bind in more conserved exons and to amplify (if possible) more variable intron regions (Meyer and Salzburger, 2012).

4.1. Single gene-tree discordance and evaluation of the strength of the phylogenetic signal

Not surprisingly, the individual single locus datasets did not contain enough phylogenetic information to accurately resolve the phylogenetic relationships among the cichlid tribes of LT. Most single locus trees were not very well resolved, the branch support values in these trees were generally rather low, and all 42 single locus topologies differed at least to some extent (in part because of the occurrence of polytomies; not shown). Overall, however, many of the single locus topologies follow a general trend as is illustrated in the average consensus network shown in Fig. 4. Many branches, and especially the monophyly of cichlid tribes, are well supported across the datasets. However, the consensus network indicates certain areas of uncertainties, which might result from hybridization and/or incomplete lineage sorting or simply reflect the low power of resolution in some of the individual markers (see below).

In order to estimate the strength of the phylogenetic signal as a function of dataset size and to evaluate whether our dataset contained enough phylogenetic information, we applied a strategy that compares tree topologies inferred from randomly chosen datasets with varying numbers of markers per alignment on the basis of their *K*-scores (Camargo et al., 2012). More specifically, we compiled datasets from 1 to 41 randomly chosen markers (in 20 replications each) and compared the ML trees based on these marker sets to the tree produced in the same way from the full dataset containing all 42 concatenated markers. Obviously, and as expected, the topologies resulting from the randomly drawn marker sets become increasingly similar to the best tree obtained with 42 markers the more markers are included in each concatenated dataset (Fig. 6a). Also, differences between equally large and mutually exclusive marker sets generally decrease with increases in the number of markers included in both sets (Fig. 6b). The same decrease was observed when trees were produced from two sets of markers that were identified as topologically concordant within each set, but discordant between sets (Fig. 6c). However, topological differences were generally slightly lower when marker sets were discordant to each other (Fig. 6c). This was unexpected but could in part be explained if the phylogenetic histories of marker sets 3–6 (which are included in Fig. 6b, but excluded from Fig. 6c) are even more discordant than those of marker sets 1 and 2.

Importantly, while all tree topologies resulting from datasets of 1–5 markers were significantly distinct from the best tree according to both SH and AU tests, inferred trees become successively more similar with an increasing number of markers, and statistically indifferent from the best tree when more than 34 markers are included (light blue lines in Fig. 6a). These results suggest that our full dataset is large enough to reliably resolve the phylogenetic history of the LT cichlid fishes. Whether or not an extension of our marker set to even more than 42 markers would provide additional phylogenetic signal remains to be tested.

4.2. A threefold strategy for phylogenetic analyses in LT cichlids

In order to account for potential problems with dataset concatenation (see below), we opted to apply three strategies to analyze our data. In a first step, we performed ML and BI phylogenetic analyses with a concatenated dataset containing all 42 markers of all 45 species. These analyses were based of the naïve assumptions that all gene histories equally reflect the species tree, and that the 'true' phylogenetic signal should dominate over phylogenetic noise in a large enough dataset (Rokas et al., 2003). The usage of the concatenated dataset is further backed up by our phylogenetic analyses of randomly chosen subsets of varying numbers of

markers, which demonstrate that the phylogenetic signal improves with increasing number of included markers (Fig. 6).

Although concatenation of multiple markers is often thought to improve accuracy (Bayzid and Warnow, 2013; Chen and Li, 2001; Rokas et al., 2003; but see Salichos and Rokas, 2013), this approach assumes that genes share a common evolutionary history, and it has been shown that violation of this assumption can lead to strongly supported yet incorrect phylogenies (Degnan and Rosenberg, 2009; Gadagkar et al., 2005; Kubatko and Degnan, 2007; Salichos and Rokas, 2013). One situation, in which concatenation may lead to inconsistent species tree estimates, is incomplete lineage sorting (Degnan and Rosenberg, 2009; Kubatko and Degnan, 2007; Yang and Rannala, 2012). We thus, in a second approach, applied a gene tree discordance test with CONCATERPILLAR to evaluate the incongruence between individual gene trees. This test suggested the existence of six sets of markers that were concordant within them, but discordant between each other. The two largest sets, containing 14 and 13 markers respectively, were then subjected to in-depth phylogenetic analysis.

As a third strategy, we performed a Bayesian concordance analysis with BUCKy, which accounts for uncertainty and variability in the individual locus phylogenies and has been shown to deal well with incomplete lineage sorting (Chung and Ané, 2011; Knowles and Kubatko, 2011; Yang and Warnow, 2011). In this analysis, we pruned our dataset to one species per tribe.

Overall, the three strategies applied to analyze our multi-marker dataset resulted in congruent topologies. All analyses confirm the monophyly of the LT tribes (in cases where more than one representative was included; this does, hence, not apply to the BUCKy analysis with the reduced taxon set). In all analyses, the Tylochromini, Oreochromini and Tilapiini were resolved outside of all other included species. The representatives of the Trematocarini and the Bathybatini always formed a clade, and were, together with *B. microlepis* (Boulengerochromini), consistently placed as sister-group to the remaining cichlid tribes; the Cyprichromini and Perissodini always clustered together. Furthermore, in all analyses except in those based on subset 1 of CONCATERPILLAR, the Lamprologini were resolved as sister group to the 'H-linage' consisting of Cyphotilapiini, Limnochromini, Cyprichromini, Perissodini, Ectodini, Eretmodini and Haplochromini. In all analyses, the Eretmodini appear as a member of the 'H-lineage' and, with one exception (i.e. subset 2 of CONCATERPILLAR), appear as sister-group to the Haplochromini.

Within the 'H-lineage', the relationships of the cichlid tribes differed between the three approaches. Especially the analysis of subset 1 of CONCATERPILLAR revealed a rather different topology, whereas in subset 2 the relative position of the Eretmodini and Ectodini varied in comparison to the other approaches. Note, however, that the two largest subsets of markers identified by CONCATERPILLAR contain only 14 (subset 1) and 13 markers (subset 2), respectively. Our analyses have shown that sets with as many as 34 markers can still produce significantly different trees for the same set of taxa. The phylogenetic hypotheses resulting from these small marker sets (Fig. 2a and b) should thus be taken with caution.

Taken together, we believe that, in our case, the concatenation of all markers is a justified strategy (Fig. 3), as it leads to the best-supported tree topologies, which are backed-up by similar results in both the average consensus network (Fig. 4) and the Bayesian concordance analysis (Fig. 5). The concatenation strategy is further supported by our phylogenetic signal tests, which show that the largest datasets lead to significantly more robust topologies (Fig. 6), whereas the subsets suggested by CONCATERPILLAR may not contain enough phylogenetic information. At the same time, these tests indicate the presence of a sufficient phylogenetic signal in the concatenated dataset, so that remaining uncertainties in the resultant tree topologies (GARLI, RAxML and MrBayes analyses of concatenated dataset and subsets) should not be due to lacking power of resolution ('soft polytomy' problem). Instead, it appears that the remaining uncertainties in our trees, most notably the phylogenetic relationships among 'H-lineage' tribes (see Figs. 2–4), are due to high speciation rates at the onset of radiation of the LT mouthbrooders ('hard polytomy' problem), past events of hybridization, and/or the persistence of ancestral polymorphisms. It has previously been recognized that it is notoriously difficult to resolve, with the available methodology, the phylogenetic relationships among lineages that emerged from adaptive radiation events (Glor, 2010), which is not least due to the fact that such tree topologies are expected to be 'bottom-heavy' (Gavrilets and Vose, 2005).

4.3. Conclusions

With this study, we present a novel hypothesis for the phylogenetic relationships among East African cichlid tribes, which is based on the largest set of ncDNA sequences so far, and which differs from all previous hypotheses (Fig. 1). Our analyses provide strong support for the monophyly of LT mouthbrooding cichlids (i.e. the 'H-lineage' of Nishida, 1991) as sister-group to the substrate spawning Lamprologini. We thus confirm the scenario that both lineages have radiated in parallel within LT (Salzburger and Meyer, 2004), leading to some intriguing cases of convergent evolution (Muschick et al., 2012). The clustering of the tribes within the 'H-lineage' generally reflects the life styles and habitat use of the respective tribes. The Cyprichromini and Perissodini, which are consistently put together (Figs. 2–5), are both adapted to the open-water column; the Cyphotilapiini and Limnochromini, which cluster together in most analyses (Figs. 2–5, excluding 2A), are restricted to deep-water habitats; and the Ectodini, Eretmodini and Haplochromini dominate (together with many lamprologine species) the shallow waters of LT. Our phylogenies thus reveal the general trend that the less species-rich cichlid tribes in LT (including the Bathybatini, Boulengerochromini and Trematocarini) occupy less-productive habitats such as the open-water column or deeper areas, whereas the generally more species-rich tribes of the 'H-lineage' dominate the more-productive and generally preferred shallow/rocky habitats (Muschick et al., 2012).

We further postulate a nested position of the Eretmodini within the 'H-lineage', as sister-group to the Haplochromini, which is in clear contrast to most of the studies relying on mtDNA markers (Clabaut et al., 2005; Day et al., 2008; Kocher et al., 1995; Muschick et al., 2012), yet in concordance to allozyme data (Nishida, 1991) and ncDNA phylogenies (Friedman et al., 2013). The obvious discordance between the Lamprologini-like mtDNA and Haplochromini-like ncDNA in the Eretmodini can either be explained by incomplete mtDNA lineage sorting, or, more likely, by an ancient hybridization event (Meng and Kubatko, 2009). The positions of the oldest tribes (Tylochromini, Oreochromini, Trematocarini, Bathybatini, Boulengerochromini) are largely in agreement with previous studies, as most studies suggested a sister-group relationship between the Bathybatini and Trematocarini (Clabaut et al., 2005; Day et al., 2008; Salzburger et al., 2002a) and placed the Oreochromini outside of this group (Friedman et al., 2013; Muschick et al., 2012; Salzburger et al., 2002a). The placement of the Boulengerochromini differed slightly between our analyses, but in all cases this monotypic tribe was resolved outside the clade formed by the Lamprologini and the 'H-lineage'.

5. Outlook

With this study, we provide a strong phylogenetic hypothesis for the cichlid tribes in LT based on 42 ncDNA makers. Yet, we also identified remaining areas of uncertainties, especially with respect

to the phylogenetic relationships of the mouthbrooding tribes within the 'H-lineage'. Future analyses should focus on the amount and relative proportion of shared genes among the different cichlid lineages to allow further insights into stochastic processes such as incomplete lineage sorting or hybridization. To this end, we recommend the usage of much larger datasets such as whole transcriptomes or genomes. RAD-sequencing could also provide a large random sample of ncDNA loci, although the current read lengths render the phylogenetic inference based on individual loci problematic. Another important next step to understand the evolutionary history of LT cichlids and to establish a species tree would be to perform coalescent-based analysis with BEST and *BEAST (Liu, 2008; Heled and Drummond, 2010), using phased alleles and more individuals per species. Finally, future analyses should increase taxon sampling, ultimately leading to a complete species tree for the cichlid species of LT.

Data accessibility

All sequences are accessible in Genbank KP129679-KP131427 and KM263618-KM263752.

Trees are deposited at Treebase (http://purl.org/phylo/treebase/phylows/study/TB2:S16660).

Acknowledgments

We thank Moritz Muschick, Adrian Indermaur and Frauke Münzel for help with sampling, and Ole Seehausen for providing tissues of Lake Victoria cichlids. We are grateful to Brigitte Aeschbach, Nicolas Boileau as well as Astrid Böhne and Emilía M. Santos for assistance in the laboratory, and to Christof Wunderlin and Georges Wigger from Microsynth for help with library preparation and 454 sequencing. We thank Stuart Willis and Cecile Ané for discussing the BUCKy analysis, and Sebastian Höhna for help with MrBayes. Finally, we would like to thank the editor Guillermo Orti, two anonymous reviewers, Heinz Büscher, Adrian Indermaur, and Uli Schliewen for discussion and helpful suggestions on the manuscript. This study was supported by the University of Basel, the Freiwillige Akademische Gesellschaft (FAG) Basel and the Burckhardt-Bürgin Stiftung (to BSM), the Swiss National Science Foundation (Grant PBBSP3-138680 to MM and Grants 3100A0_122458 and 3100A0_138224 to WS) and the European Research Council (ERC; Starting Grant "INTERGENADAPT" and Consolidator Grant "Cichlid∼X" to WS).

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.ympev.2014.10.009.

References

Albertson, R.C., Streelman, J.T., Kocher, T.D., 2003. Directional selection has shaped the oral jaws of Lake Malawi cichlid fishes. Proc. Natl. Acad. Sci. USA 100, 5252–5257.

Ane, C., Larget, B., Baum, D.A., Smith, S.D., Rokas, A., 2007. Bayesian estimation of concordance among gene trees. Mol. Biol. Evol. 24, 412–426.

Ballard, J.W.O., Whitlock, M.C., 2004. The incomplete natural history of mitochondria. Mol. Ecol. 13, 729–744.

Baum, D., 2007. Concordance trees, concordance factors, and the exploration of reticulate genealogy. Taxon 56, 417–426.

Bayzid, M.S., Warnow, T., 2013. Naive binning improves phylogenomic analyses. Bioinformatics 29, 2277–2284.

Brawand, D. et al., 2014. The genomic substrate for adaptive radiation in African cichlid fish. Nature 513, 375–381.

Camargo, A., Avila, L.J., Morando, M., Sites, J.W., 2012. Accuracy and precision of species trees: effects of locus, individual, and base pair sampling on inference of species trees in lizards of the Liolaemus darwinii group (Squamata, Liolaemidae). Syst. Biol. 61, 272–288.

Chen, F.-C., Li, W.-H., 2001. Genomic divergences between humans and other hominoids and the effective population size of the common ancestor of humans and chimpanzees. Am. J. Human Genet. 68, 444–456.

Chow, S., Hazama, K., 1998. Universal PCR primers for S7 ribosomal protein gene introns in fish. Mol. Ecol. 7, 1255–1256.

Chung, Y., Ané, C., 2011. Comparing two bayesian methods for gene tree/species tree reconstruction: simulations with incomplete lineage sorting and horizontal gene transfer. Syst. Biol. 60, 261–275.

Clabaut, C., Salzburger, W., Meyer, A., 2005. Comparative phylogenetic analyses of the adaptive radiation of Lake Tanganyika cichlid fish: nuclear sequences are less homoplasious but also less informative than mitochondrial DNA. J. Mol. Evol. 61, 666–681.

Cohen, A.S., Lezzar, K.E., Tiercelin, J.J., Soreghan, M., 1997. New palaeogeographic and lake-level reconstructions of Lake Tanganyika: implications for tectonic, climatic and biological evolution in a rift lake. Basin Res. 9, 107–132.

Day, J.J., Cotton, J.A., Barraclough, T.G., 2008. Tempo and mode of diversification of Lake Tanganyika cichlid fishes. PLoS ONE 3, e1730.

Degnan, J.H., Rosenberg, N.A., 2009. Gene tree discordance, phylogenetic inference and the multispecies coalescent. Trends Ecol. Evol. 24, 332–340.

Duftner, N., Koblmuller, S., Sturmbauer, C., 2005. Evolutionary relationships of the Limnochromini, a tribe of benthic deepwater cichlid fish endemic to Lake Tanganyika, East Africa. J. Mol. Evol. 60, 277–289.

Dunz, A.R., Schliewen, U.K., 2013. Molecular phylogeny and revised classification of the haplotilapiine cichlid fishes formerly referred to as "*Tilapia*". Mol. Phylogenet. Evol. 68, 64–80.

Friedman, M., Keck, B.P., Dornburg, A., Eytan, R.I., Martin, C.H., Hulsey, C.D., Wainwright, P.C., Near, T.J., 2013. Molecular and fossil evidence place the origin of cichlid fishes long after Gondwanan rifting. Proc. R. Soc. Lond. B 280, 20131733.

Gadagkar, S.R., Rosenberg, M.S., Kumar, S., 2005. Inferring species phylogenies from multiple genes: concatenated sequence tree versus consensus gene tree. J. Exp. Zool. B. Mol. Dev. Evol. 304, 64–74.

Gavrilets, S., Vose, A., 2005. Dynamic patterns of adaptive radiation. Proc. Natl. Acad. Sci. USA 102, 18040–18045.

Genner, M.J., Seehausen, O., Lunt, D.H., Joyce, D.A., Shaw, P.W., Carvalho, G.R., Turner, G.F., 2007. Age of cichlids: new dates for ancient lake fish radiations. Mol. Biol. Evol. 24, 1269–1282.

Glor, R.E., 2010. Phylogenetic insights on adaptive radiation. Ann. Rev. Ecol. Evol. Syst. 41, 251–270.

Heled, J., Drummond, A.J., 2010. Bayesian inference of species trees from multilocus data. Mol. Biol. Evol. 27, 570–580.

Huson, D.H., Bryant, D., 2006. Application of phylogenetic networks in evolutionary studies. Mol. Biol. Evol. 23, 254–267.

Joyce, D.A., Lunt, D.H., Genner, M.J., Turner, G.F., Bills, R., Seehausen, O., 2011. Repeated colonization and hybridization in Lake Malawi cichlids. Curr. Biol. 21, R108–109.

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

Kirchberger, P.C., Sefc, K.M., Sturmbauer, C., Koblmüller, S., 2014. Outgroup effects on root position and tree topology in the AFLP phylogeny of a rapidly radiating lineage of cichlid fish. Mol. Phylogenet. Evol. 70, 57–62.

Klett, V., Meyer, A., 2002. What, if anything, is a Tilapia? – mitochondrial ND2 phylogeny of tilapiines and the evolution of parental care systems in the African cichlid fishes. Mol. Biol. Evol. 19, 865–883.

Knowles, L.L., Kubatko, L.S., 2011. Estimating Speciest: Practical and Theoretical Aspects. Wiley.

Koblmüller, S., Duftner, N., Katongo, C., Phiri, H., Sturmbauer, C., 2005. Ancient divergence in bathypelagic Lake Tanganyika deepwater cichlids: mitochondrial phylogeny of the tribe Bathybatini. J. Mol. Evol. 60, 297–314.

Koblmüller, S., Duftner, N., Sefc, K.M., Aibara, M., Stipacek, M., Blanc, M., Egger, B., Sturmbauer, C., 2007. Reticulate phylogeny of gastropod-shell-breeding cichlids from Lake Tanganyika – the result of repeated introgressive hybridization. BMC Evol. Biol. 7, 7.

Koblmüller, S., Schliewen, U.K., Duftner, N., Sefc, K.M., Katongo, C., Sturmbauer, C., 2008a. Age and spread of the haplochromine cichlid fishes in Africa. Mol. Phylogenet. Evol. 49, 153–169.

Koblmüller, S., Sefc, K.M., Sturmbauer, C., 2008b. The Lake Tanganyika cichlid species assemblage: recent advances in molecular phylogenetics. Hydrobiol. 615, 5–20.

Koch, M., Koblmüller, S., Sefc, K., 2007. Evolutionary history of the endemic Lake Tanganyika cichlid fish *Tylochromis polylepis*: a recent intruder to a mature adaptive radiation. J. Zool. Syst. Evol. Res. 45, 64–71.

Kocher, T.D., 2004. Adaptive evolution and explosive speciation: the cichlid fish model. Nature Rev. Genet. 5, 288–298.

Kocher, T.D., Conroy, J.A., McKaye, K.R., Stauffer, J.R., Lockwood, S.F., 1995. Evolution of NADH dehydrogenase subunit 2 in east African cichlid fish. Mol. Phylogenet. Evol. 4, 420–432.

Kubatko, L., Degnan, J., 2007. Inconsistency of phylogenetic estimates from concatenated data under coalescence. Syst. Biol. 56, 17–24.

Kumar, S., Skjaeveland, A., Orr, R.J., Enger, P., Ruden, T., Mevik, B.H., Burki, F., Botnen, A., Shalchian-Tabrizi, K., 2009. AIR: a batch-oriented web program package for construction of supermatrices ready for phylogenomic analyses. BMC Bioinf. 10, 357.

Lanfear, R., Calcott, B., Ho, S.Y., Guindon, S., 2012. Partitionfinder: combined selection of partitioning schemes and substitution models for phylogenetic analyses. Mol. Biol. Evol. 29, 1695–1701.
Lang, M., Miyake, T., Braasch, I., Tinnemore, D., Siegel, N., Salzburger, W., Amemiya, C.T., Meyer, A., 2006. A BAC library of the East African haplochromine cichlid fish *Astatotilapia burtoni*. J. Exp. Zool. B. Mol. Dev. Evol. 306, 35–44.
Larget, B.R., Kotha, S.K., Dewey, C.N., Ane, C., 2010. BUCKy: gene tree/species tree reconciliation with Bayesian concordance analysis. Bioinformatics 26, 2910–2911.
Leigh, J.W., Susko, E., Baumgartner, M., Roger, A.J., 2008. Testing congruence in phylogenomic analysis. Syst. Biol. 57, 104–115.
Lessa, E.P., 1992. Rapid surveying of DNA sequence variation in natural populations. Mol. Biol. Evol. 9, 323.
Lewis, P.O., 2001. A likelihood approach to estimating phylogeny from discrete morphological character data. Syst. Biol. 50, 913–925.
Li, H., Durbin, R., 2010. Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics 26, 589–595.
Li, C., Ortí, G., Zhang, G., Lu, G., 2007. A practical approach to phylogenomics: the phylogeny of ray-finned fish (Actinopterygii) as a case study. BMC Evol. Biol. 7, 44.
Liu, L., 2008. BEST: Bayesian estimation of species trees under the coalescent model. Bioinformatics 24, 2542–2543.
Maddison, W., 1989. Reconstructing character evolution on polytomous cladograms. Cladistics 5, 365–377.
McCormack, J.E., Hird, S.M., Zellmer, A.J., Carstens, B.C., Brumfield, R.T., 2013. Applications of next-generation sequencing to phylogeography and phylogenetics. Mol. Phylogenet. Evol. 66, 526–538.
Meng, C., Kubatko, L.S., 2009. Detecting hybrid speciation in the presence of incomplete lineage sorting using gene tree incongruence: a model. Theor. Popul. Biol. 75, 35–45.
Meyer, B.S., Salzburger, W., 2012. A novel primer set for multilocus phylogenetic inference in East African cichlid fishes. Mol. Ecol. Res. 12, 1097–1104.
Miller, M.A., Pfeiffer, W., Schwartz, T., 2010. Creating the CIPRES Science Gateway for inference of large phylogenetic trees. Gateway Comput. Environ. Workshop (GCE) 2010, 1–8.
Müller, K., 2005. SeqState: primer design and sequence statistics for phylogenetic DNA datasets. Appl. Bioinf. 4, 65–69.
Muschick, M., Indermaur, A., Salzburger, W., 2012. Convergent evolution within an adaptive radiation of cichlid fishes. Curr. Biol. 22, 2362–2368.
Nishida, M., 1991. Lake Tanganyika as an evolutionary reservoir of old lineages of East African cichlid fishes: inferences from allozyme data. Experientia 47, 974–979.
Nylander, J.A.A., Wilgenbusch, J.C., Warren, D.L., Swofford, D.L., 2008. AWTY (are we there yet?): a system for graphical exploration of MCMC convergence in Bayesian phylogenetics. Bioinformatics 24, 581–583.
Pamilo, P., Nei, M., 1988. Relationships between gene trees and species trees. Mol. Biol. Evol. 5, 568–583.
Poll, M., 1986. Classification des Cichlidae du lac Tanganika. Tribus, genres et especes. Academie Royale de Belgique. Classe des Sciences. Memoires 45, 1–163.
Rambaut, A., Drummond, A.J., 2007. Tracer v1.5 <http://beast.bio.ed.ac.uk/Tracer>.
Rokas, A., Williams, B.L., King, N., Carroll, S.B., 2003. Genome-scale approaches to resolving incongruence in molecular phylogenies. Nature 425, 798–804.
Ronquist, F., Teslenko, M., van der Mark, P., Ayres, D.L., Darling, A., Hohna, S., Larget, B., Liu, L., Suchard, M.A., Huelsenbeck, J.P., 2012. MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. Syst. Biol. 61, 539–542.
Salichos, L., Rokas, A., 2013. Inferring ancient divergences requires genes with strong phylogenetic signals. Nature 497, 327–331.
Salzburger, W., 2009. The interaction of sexually and naturally selected traits in the adaptive radiations of cichlid fishes. Mol. Ecol. 18, 169–185.
Salzburger, W., Meyer, A., 2004. The species flocks of East African cichlid fishes: recent advances in molecular phylogenetics and population genetics. Naturwissenschaften 91, 277–290.
Salzburger, W., Meyer, A., Baric, S., Verheyen, E., Sturmbauer, C., 2002a. Phylogeny of the Lake Tanganyika cichlid species flock and its relationship to the Central and East African haplochromine cichlid fish faunas. Syst. Biol. 51, 113–135.
Salzburger, W., Baric, S., Sturmbauer, C., 2002b. Speciation via introgressive hybridization in East African cichlids? Mol. Ecol. 11, 619–625.
Salzburger, W., Mack, T., Verheyen, E., Meyer, A., 2005. Out of Tanganyika: genesis, explosive speciation, key-innovations and phylogeography of the haplochromine cichlid fishes. BMC Evol. Biol. 5, 17.
Salzburger, W., Van Bocxlaer, B., Cohen, A.S., 2014. Ecology and evolution of the African Great Lakes and their faunas. Annu. Rev. Ecol. Evol. Syst. 45, 519–545.
Santos, M.E., Salzburger, W., 2012. Evolution. How cichlids diversify. Science 338, 619–621.
Santos, M.E., Braasch, I., Boileau, N., Meyer, B.S., Sauteur, L., Böhne, A., Belting, H.-G., Affolter, M., Salzburger, W., 2014. The evolution of cichlid fish egg-spots is linked with a cis-regulatory change. Nat. Commun. 5. http://dx.doi.org/10.1038/ncomms6149.
Schelly, R.C., Stiassny, M.L.J., 2004. Revision of the Congo River *Lamprologus* Schilthuis, 1891 (Teleostei: Cichlidae), with descriptions of two new species. Am. Mus. Nov. 3451, 1–40.
Schelly, R., Stiassny, M.L., Seegers, L., 2003. *Neolamprologus devosi* sp. n., a new riverine lamprologine cichlid (Teleostei, Cichlidae) from the lower Malagarasi River. Tanzania. Zootaxa 373, 1–11.
Schmieder, R., Edwards, R., 2011. Quality control and preprocessing of metagenomic datasets. Bioinformatics 27, 863–864.
Schwarz, G., 1978. Estimating the dimension of a model. Ann. Stat. 6, 461–464.
Schwarzer, J., Swartz, E.R., Vreven, E., Snoeks, J., Cotterill, F.P.D., Misof, B., Schliewen, U.K., 2012. Repeated trans-watershed hybridization among haplochromine cichlids (Cichlidae) was triggered by Neogene landscape evolution. Proc. R. Soc. Lond. B 279, 4389–4398.
Seehausen, O., 2004. Hybridization and adaptive radiation. Trends Ecol. Evol. 19, 198–207.
Seehausen, O., 2006. African cichlid fish: a model system in adaptive radiation research. Proc. R. Soc. Lond. B 273, 1987–1998.
Shimodaira, H., 2002. An approximately unbiased test of phylogenetic tree selection. Syst. Biol. 51, 492–508.
Shimodaira, H., Hasegawa, M., 1999. Multiple comparisons of log-likelihoods with applications to phylogenetic inference. Mol. Biol. Evol. 16, 1114–1116.
Simmons, M.P., Ochoterena, H., 2000. Gaps as characters in sequence-based phylogenetic analyses. Syst. Biol. 49, 369–381.
Slade, R.W., Moritz, C., Heideman, A., Hale, P.T., 1993. Rapid assessment of single-copy nuclear DNA variation in diverse species. Mol. Ecol. 2, 359–373.
Slowinski, J.B., 2001. Molecular polytomies. Mol. Phylogenet. Evol. 19, 114–120.
Snoeks, J., 2000. How well known is the ichthyodiversity of the large East African lakes? In: Rossiter, A., Kawanabe, H. (Eds.), Advances in Ecological Research, 0065-2504, 31. Academic Press, ISBN 9780120139316, pp. 17–38. http://dx.doi.org/10.1016/S0065-2504(00)31005-4.
Soria-Carrasco, V., Talavera, G., Igea, J., Castresana, J., 2007. The K tree score: quantification of differences in the relative branch length and topology of phylogenetic trees. Bioinformatics 23, 2954–2956.
Stamatakis, A., 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics 22, 2688–2690.
Stiassny, M., 1990. Tylochromis, relationships and the phylogenetic status of the African Cichlidae. Am. Mus. Nov. 2993, 1–14.
Sturmbauer, C., Meyer, A., 1993. Mitochondrial phylogeny of the endemic mouthbrooding lineages of cichlid fishes from Lake Tanganyika in eastern Africa. Mol. Biol. Evol. 10, 751–768.
Sturmbauer, C., Hainz, U., Baric, S., Verheyen, E., Salzburger, W., 2003. Evolution of the tribe Tropheini from Lake Tanganyika: synchronized explosive speciation producing multiple evolutionary parallelism. Hydrobiologia 500, 51–64.
Sturmbauer, C., Salzburger, W., Duftner, N., Schelly, R., Koblmuller, S., 2010. Evolutionary history of the Lake Tanganyika cichlid tribe Lamprologini (Teleostei: Perciformes) derived from mitochondrial and nuclear DNA data. Mol. Phylogenet. Evol. 57, 266–284.
Sukumaran, J., Holder, M.T., 2010. DendroPy: a Python library for phylogenetic computing. Bioinformatics 26, 1569–1571.
Swofford, D.L., 2003. PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods). Sinauer Associates, Sunderland, Massachusetts.
Takahashi, T., 2003. Systematics of Tanganyikan cichlid fishes (Teleostei: Perciformes). Ichthyol. Res. 50, 367–382.
Takahashi, K., Terai, Y., Nishida, M., Okada, N., 2001. Phylogenetic relationships and ancient incomplete lineage sorting among cichlid fishes in Lake Tanganyika as revealed by analysis of the insertion of retroposons. Mol. Biol. Evol. 18, 2057–2066.
Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M., Kumar, S., 2011. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol. Biol. Evol. 28, 2731–2739.
Turner, G.F., Seehausen, O., Knight, M.E., Allender, C.J., Robinson, R.L., 2001. How many species of cichlid fishes are there in African lakes? Mol. Ecol. 10, 793–806.
Verheyen, E., Salzburger, W., Snoeks, J., Meyer, A., 2003. Origin of the superflock of cichlid fishes from Lake Victoria, East Africa. Science.
Wagner, C.E., Harmon, L.J., Seehausen, O., 2012. Ecological opportunity and sexual selection together predict adaptive radiation. Nature 487, 366–369.
Walsh, H.E., Kidd, M.G., Moum, T., Friesen, V.L., 1999. Polytomies and the power of phylogenetic inference. Evolution 53, 932–937.
Whitfield, J.B., Lockhart, P.J., 2007. Deciphering ancient rapid radiations. Trends Ecol. Evol. 22, 258–265.
Willis, S.C., Farias, I.P., Orti, G., 2013. Multi-locus species tree for the Amazonian peacock basses (Cichlidae: *Cichla*): Emergent phylogenetic signal despite limited nuclear variation. Mol. Phylogenet. Evol. 69, 479–490.
Won, Y.-J., Sivasundar, A., Wang, Y., Hey, J., 2005. On the origin of Lake Malawi cichlid species: a population genetic analysis of divergence. Proc. Natl. Acad. Sci. USA 102 (Suppl. 1), 6581–6586.
Yang, Z., Rannala, B., 2012. Molecular phylogenetics: principles and practice. Nature Rev. Genet., 303–314.
Yang, J., Warnow, T., 2011. Fast and accurate methods for phylogenomic analyses. BMC Bioinf. 12 (Suppl 9), S4.
Zwickl, D.J., 2006. Genetic Algorithm Approaches for the Phylogenetic Analysis of Large Biological Sequence Datasets Under the Maximum Likelihood Criterion. PhD Thesis. The University of Texas at Austin.