A novel primer set for multilocus phylogenetic inference in East African cichlid fishes

BRITTA S. MEYER and WALTER SALZBURGER
Zoological Institute, University of Basel, Vesalgasse 1, 4051 Basel, Switzerland

Abstract

The cichlid fishes in the East African Great Lakes are a prime model system for the study of adaptive radiation. Therefore, the availability of an elaborate phylogenetic framework is an important prerequisite. Previous phylogenetic hypotheses on East African cichlids are mainly based on mitochondrial and/or fragment-based markers, and, to date, no taxon-rich phylogeny exists that is based on multilocus DNA sequence data. Here, we present the design of an extensive new primer set (24 nuclear markers) for East African cichlids that will be used for multilocus phylogenetic analyses in the future. The primers are designed to work for both Sanger sequencing and next-generation sequencing with the 454 technology. A saprozoa is of principle, we validate these primers in a phylogenetically representative set of 16 cichlid species from Lake Tanganyika and main river systems in the area and provide a basic evaluation of the markers with respect to marker length and diversity indices.

Keywords: adaptive radiation, cichlid species flocks, nuclear markers, organismal diversification

Received 20 December 2011; revision received 8 June 2012; accepted 12 June 2012

Introduction

Cichlid fishes (Cichlidae) belong to one of the most species-rich families of vertebrates, with a distribution range from Africa including Madagascar, to Central and South America and South India. The Great Lakes in the East African Rift Valley harbour the largest and most diverse species flocks of cichlid fishes (Snoeks 2000; Turner et al. 2001) and are regarded as prime model systems to study evolutionary processes (reviewed in Kocher 2004; Salzburger 2009; Seehausen 2006). Among the three main species flocks, that of Lake Victoria, Lake Malawi and Lake Tanganyika, the latter is the morphologically, behaviourally, ecologically and genetically most diverse (Sturmbauer & Meyer 1992; Salzburger et al. 2002b; Young et al. 2009). This is due to the greater age of the lake, estimated between nine and 12 Ma (Cohen et al. 1993), and, consequently, the greater age of the radiation itself (Gnner et al. 2007; Koblmüller et al. 2008; Schwarzer et al. 2009). Moreover, and unlike the species flocks of Lake Victoria and Lake Malawi, which exclusively consist of species of the haplochrome sublineage (‘tribe’), the cichlid assemblage in Lake Tanganyika consists of 12–16 tribes, of which the haplochromines are but one (Poll 1986; Salzburger et al. 2002b; Takahashi 2003; Clabaut et al. 2005).

Since their discovery at the turn of the nineteenth century, the species flocks of cichlids in East Africa have been in the centre of empirical and theoretical research. The main focus has always been on speciation in general, and in particular, on the tempo and mode of diversification, the possible triggers and the progress of adaptive radiations, the respective role of sexual and natural selection and the role of evolutionary key innovations (Verheyen et al. 2003; Salzburger et al. 2005, 2007; Day et al. 2008; Seehausen et al. 2008; Salzburger 2009). Importantly, most of this research depends on phylogenetic hypotheses, which appear difficult to obtain in the rapidly evolving assemblages of cichlids in East Africa (Kocher 2003). Especially in the comparably young cichlid radiations of lakes Malawi and Victoria, there is only limited genetic variation in mitochondrial markers between both species and genera, and haplotype sharing is a common phenomenon (Meyer et al. 1990; Parker & Kornfield 1997; Shaw et al. 2000; Verheyen et al. 2003). The fragment-based amplified fragment length polymorphism (AFLP) method provided better resolution here (Albertson et al. 1999; Allender et al. 2003; Joyce et al. 2011), although comprehensive phylogenies are still lacking for cichlids from lakes Malawi and Victoria.

A more extensive phylogenetic framework is available for the cichlid species flock of Lake Tanganyika, which also includes analyses of its sublineages (‘tribes’). Most of the available phylogenetic hypotheses are based on
mitochondrial markers (e.g. Cyprichromini: Brandstätter et al. 2005; Lamprologini: Day et al. 2007; Limnochromini: Duftner et al. 2005; Bathybatini: Koblmüller et al. 2005; Ectodini: Koblmüller et al. 2004; Haplochromini: Salzburger et al. 2005; Koblmüller et al. 2008). Fewer studies used a combination of sequence-based nuclear and mitochondrial markers (Salzburger et al. 2002a; Clabaut et al. 2005; Schelly et al. 2006; Nevado et al. 2009) or AFLPs and mitochondrial markers (Egger et al. 2007; Koblmüller et al. 2007a,b, 2010; Takahashi et al. 2007; Sturmbauer et al. 2010). These studies often led to new insights regarding hybridization, introgression or incomplete lineage sorting events (Nevado et al. 2009, 2011; Koblmüller et al. 2010). However, no taxon-rich phylogenetic study exists that is based on sequence data from various nuclear markers. This is in contrast to the many advantages that a (nuclear) multilocus phylogeny would provide. Most importantly, a species tree inferred from the gene trees of many independent loci should be more accurate than a species tree obtained from a few loci or a single locus only (Pamilo & Nei 1988).

Here, we present the design and general validation of primer pairs for 24 nuclear loci in East African cichlids. Our main goal was to obtain a set of nuclear markers for multilocus phylogenetic analyses. We focused on the development of markers with a length suitable for high-throughput sequencing. At the same time, we designed primers to amplify genes with known functions and from different functional categories. As a proof of principle, we tested our marker set in 16 East African cichlid species across a broad phylogenetic range and performed a phylogenetic analysis.

Materials and methods

Primer design

First, we defined the following general requirements for our primer sets:

1. the primers should work in a phylogenetically representative set of East African cichlid fishes
2. the primers should amplify between ca. 400–600 bp (based on the current read length of 454 sequencing/GS FLX Titanium)
3. they should have a maximal length of 24 bp (based on recommendations for fusion primer design)
4. all primers should have a similar melting temperature (Tm) at an optimum between 57–59 °C (according to the table of thermodynamic parameters from (SantaLucia (1998))
5. the genes to be amplified should be well characterized.

We first screened the literature for candidate primers, which were then, if necessary, modified to match the above requirements. Second, to generate new markers, we selected a set of candidate genes with known functions, for example, in coloration and pigmentation, growth factor activity, (craniofacial) bone development, protein processing, cell cycle, metabolism, or as transcription factors and ribosomal proteins. In the absence of a cichlid genome assembly (at the time the study was performed), the distribution of these candidate genes across fish genomes was determined using the available assemblies of Zebrafish (Danio rerio) and Medaka (Oryzias latipes) in Ensembl (Flicek et al. 2011). These two assemblies, in combination with available cichlid cDNA/EST sequences (Watanabe et al. 2004; Tsai et al. 2007; Salzburger et al. 2008; Time et al. 2008; Kobayashi et al. 2009; Lee et al. 2010; Baldo et al. 2011), were also used to infer exon/intron boundaries for each locus, which was important to estimate intron lengths. The final primer design was based on additional cichlid sequences (from NCBI databases ‘wgs’ (whole genome shotgun) and ‘nr’ as well as unpublished sequences from our laboratory). To avoid the amplification of ancient paralogs, primers were designed in regions where paralogs differed. If possible, primers were designed for exon-primed intron-crossing (EPIC) markers, which anneal in conserved exons and amplify mainly the introns. Given a read length of ca. 400 bp (after trimming) by 454/GS FLX Titanium and ca. 600–800 bp by Sanger sequencing, only relatively short introns could be considered. All primers were designed with Primer-BLAST (Sayers et al. 2011), which includes the software Primer3 (Rozen & Skaletsky 2000) and a BLAST search (Altschul et al. 1990, 1997; Sayers et al. 2011), using the nr nucleotide database with the ‘taxid’ (NCBI taxonomy id) for cichlids (8113).

Taxon sampling

To assess the applicability of the newly designed primers in a broad spectrum of cichlid species, we tested them in a phylogenetically representative set of 16 cichlid species representing 12 tribes (Tylochromini, Tilapiini, Bathybatini, Eretmodini, Lamprologini, Ectodini, Cyprichromini, Perissodini, Limnochromini, Haplochromini/Tropheini, Cyphotilapiini). Tylochromis polylepis, a relatively recent colonizer of Lake Tanganyika and a representative of an ancestral lineage, was included as outgroup (Salzburger et al. 2002b; Clabaut et al. 2005; Koch et al. 2007). Note that most species are from Lake Tanganyika to account for its greater diversity in cichlid lineages; however, as we also included several haplochromines, our taxon sampling represents the entire phylogenetic spectrum of East African cichlids. Samples were collected in the years 2007 and 2008.
Molecular data

Genomic DNA was extracted from fin clips preserved in 95% ethanol, using the robotic workstation BioSprint 96 following the manufacturer’s protocol (Qiagen, Hombrechtikon, Switzerland). PCRs were performed in a final volume of 12.5 µL containing REDTaq® DNA Polymerase (0.04 units/µL), its PCR Buffer (1×) (Sigma-Aldrich, Buchs, Switzerland), 200 µM of each dNTP (Promega, Dübendorf, Switzerland), 0.2 µM of each sense and anti-sense primer (Microsynth, Balgach, Switzerland), 5–10 ng of DNA and water. The PCR conditions of all target fragments consisted of an initial denaturation for 2 min at 94 °C, followed by 32 cycles with a denaturation step at 94 °C for 30 s, an annealing step at 52–54 °C for 30 s and finalized by an extension step at 72 °C for 1 min. PCR success was evaluated using gel electrophoresis (1.5% agarose; buffered in 1·TAE). To assess the length of the PCR product, a size standard (BenchTop 100bp DNA Ladder; Promega) was added as reference to the gel. For visualization under UV-light, the gel was stained with GelRed™ (Biotium; VWR International, Dietikon, Switzerland).

Prior to DNA sequencing, the PCR products were purified from excess primers and dNTPs using ExoSAP-IT (GE Healthcare, Glattbrugg, Switzerland) following the manufacturer’s protocol. Sequencing reactions were specified in Table 1 (0.5 µL), its PCR Buffer (1×) (Sigma-Aldrich, Buchs, Switzerland), 200 µM of each dNTP (Promega, Dübendorf, Switzerland), 0.2 µM of each sense and anti-sense primer (Microsynth, Balgach, Switzerland), 5–10 ng of DNA and water. The PCR conditions of all target fragments consisted of an initial denaturation for 2 min at 94 °C, followed by 32 cycles with a denaturation step at 94 °C for 30 s, an annealing step at 52–54 °C for 30 s and finalized by an extension step at 72 °C for 1 min. PCR success was evaluated using gel electrophoresis (1.5% agarose; buffered in 1·TAE). To assess the length of the PCR product, a size standard (BenchTop 100bp DNA Ladder; Promega) was added as reference to the gel. For visualization under UV-light, the gel was stained with GelRed™ (Biotium; VWR International, Dietikon, Switzerland).

Prior to DNA sequencing, the PCR products were purified from excess primers and dNTPs using ExoSAP-IT (GE Healthcare, Glattbrugg, Switzerland) following the manufacturer’s protocol. Sequencing reactions were performed using 1.5 µL purified PCR products, the primers specified in Table 1 (0.5 µL/10 µµ), and 1 µL BigDye® Terminator v3.1 Ready Reaction Mix (Applied Biosystems, Rotkreuz, Switzerland) in a total volume of 8 µL under standard conditions (1 min 94 °C, followed by 25 cycles with 10 s at 94 °C, 20 s at 52 °C, 4 min at 60 °C). To scavenge all unincorporated BigDye® terminators, the BigDye X Terminator® Purification Kit with its standard protocol (Applied Biosystems) was used. After this purification step, sequences were obtained with the 3130x/Genetic Analyzer (Applied Biosystems). Chromatograms were edited in CODONCODE ALIGNER (CodonCode, Dedham, MA, USA). Double peaks with equally high intensities in the chromatograms were assigned as heterozygous sites (SNPs). These polymorphic sites were coded as ambiguous nucleotides following the IUPAC-IUB code.

Sequence analysis

Initial alignments were performed with MAFFT (--auto) (Katoh & Toh 2008). Thereafter, a ‘supermatrix’ was generated by concatenating the single genes of the 16 species using MESQUITE 2.73 (Maddison & Maddison 2010). We then used MEGAS (Tamura et al. 2011) to calculate genetic p-distance between the ingroup species (excluding T. polyplepis) with complete deletion, for each single gene (and, in a second step, within exons and introns separately). The percentage of missing data, the gaps and polymorphic sites were accessed using MACCLADE 4.08 (Maddison & Maddison 2005).

Phylogenetic analysis

Prior to phylogenetic analysis, we determined the best fitting substitution model for each gene with jMODELTEST 0.1.1 (Guindon & Gascuel 2003; Posada 2008) on the basis of the Bayesian information criterion (BIC) (Schwarz 1978). We first performed a maximum likelihood analysis with GARLI 2.0 (Zwickl 2006) and our partitioned supermatrix. We run ten independent replicates, which were terminated automatically after 5000 generations with no significant (P < 0.01) improvements in topology scoring. To assess confidence in the tree topology, 1000 bootstrap replicates were executed and a majority-rule consensus tree was constructed with PAUP* 4.0a114 (Swoford 2002). Bayesian phylogenetic inference for the partitioned data set was conducted with MrBayes v3.2 (Huelsenbeck & Ronquist 2001; Ronquist & Huelsenbeck 2003). The starting trees were set to be random. Prior probability distributions for all parameters were assumed to be flat. Two simultaneous MCMC were conducted for 21 000 000 generations, each of which had three heated and one cold chain, and the trees were sampled every 1000 generations. The first 25% of the sampled trees were discarded as burnin. To diagnose convergence in the two runs, we used AWTY (Nylander et al. 2008) and TRACER v1.5 (Ram-baut & Drummond 2007). The majority-rule consensus tree derived from GARLI and PAUP* as well as the tree from MrBayes were finally processed in FigTree v1.3.1 (Ram-baut 2009).

Results

We designed 24 new primer pairs that amplify nuclear markers in East African cichlid fishes. Two of these primers are variations of already existing primers. The S7 reverse primer is adopted from Chow & Hazama (1998), but with an extra degenerated nucleotide. The other primer, the bmp4 reverse primer, is a variation of an existing primer from Albertson et al. (2003), slightly elongated and with more specific nucleotides. The length of the resulting PCR products ranges between 357–707 base pairs, with an average length of 497 bp and a median of 483.5 bp. Table 1 lists all loci with their specific forward and reverse primer sequences, their location in Medaka chromosomes and the number of base pairs belonging to intron or exon.

The amplification of these loci was successful in most of the 16 tested cichlid species; on average, 15 species showed a band on the agarose gels. Sequencing success with the Sanger method was less successful, which we
attribute in part to the existence of alleles with different lengths resulting in double peaks (note: this is not an issue when using next-generation sequencing techniques). The percentage of missing data and gaps per species are listed in Table 2.

In total, we obtained a concatenated data set of 24 paralogous gene sequences containing 9669 bp. A total of 583 sites were variable (6.03%), of which 130 are parsimony-informative sites (1.3%) (calculated without the outgroup taxon *Tylochromis polyepis* and without indels or polymorphic sites coded with ambiguous IUPAC code). The combined sequence matrix consists of 5761 bp (59.58%) from exons and 3908 bp (40.42%) from introns (Table 1).

In 18 of the 24 loci, we detected heterozygous SNPs (46 SNPs in total; referred to as polymorphic sites coded with ambiguous IUPAC code). The percentage of missing data and gaps per species are listed in Table 2.

Table 1: List of the 24 primer pairs with their forward and reverse sequences using IUPAC code, the length of the respective PCR product with primers, the ENSEMBL-ID (or NCBI Accession no) of the Medaka ortholog, known chromosome position (Chr. pos.) in Medaka (M), and the amount of base pairs in exon and intron.

<table>
<thead>
<tr>
<th>Locus</th>
<th>Primer forward/reverse (5'-3')</th>
<th>PCR product (bp)</th>
<th>ID of Medaka ortholog</th>
<th>Chr pos (bp) Exon Intron</th>
</tr>
</thead>
<tbody>
<tr>
<td>bmp4</td>
<td>GAGGACCATGGCCATTTCTCGTTT/GCCAATCATATCTGCTATATTCCAGGCC</td>
<td>577</td>
<td>ENSORLG00000013304 22</td>
<td>482 0</td>
</tr>
<tr>
<td>bmp2</td>
<td>AGGCCCTTGGCAACGTAAAATTCTCGCTGCTTGGCGCAATTC</td>
<td>414</td>
<td>ENSORLG00000009772 24</td>
<td>315 0</td>
</tr>
<tr>
<td>fg6</td>
<td>CCAGAAATGTCGCTACACAGTCCGATGTCAA</td>
<td>512</td>
<td>ENSORLG00000015820 23</td>
<td>286 158</td>
</tr>
<tr>
<td>furina</td>
<td>GCTGCATGGGACAGACAGATCA/ATAGCTACCTGCGACCCGCCAAC</td>
<td>357</td>
<td>ENSORLG00000009133 3</td>
<td>154 94</td>
</tr>
<tr>
<td>runx2</td>
<td>CGGGTCGTGGTGTGAGAAGGGCCAGGCA</td>
<td>411</td>
<td>ENSORLG00000010169 24</td>
<td>95 218</td>
</tr>
<tr>
<td>shh</td>
<td>TTGCAACAAAGAAGCGCTGTCGTCGTCG</td>
<td>512</td>
<td>ENSORLG00000010463 20</td>
<td>421 0</td>
</tr>
<tr>
<td>pax9</td>
<td>TCCACCGCCTGTCATAGYGA/ACAGATGCGGAGGAAAGCCCA</td>
<td>434</td>
<td>AB187122.1 ?</td>
<td>338 0</td>
</tr>
<tr>
<td>sox10b</td>
<td>TSCRGGCTTCGGAACAATCTCAT/GTGTGCTGCGATATCTTGGCA</td>
<td>486</td>
<td>ENSORLG00000014587 8</td>
<td>310 0</td>
</tr>
<tr>
<td>ednb1</td>
<td>CTTGTCCTCTAGCCTGAT/AGGACGCCAGCACAGCAGA</td>
<td>479</td>
<td>ENSORLG00000011054 17</td>
<td>54 320</td>
</tr>
<tr>
<td>mc1r</td>
<td>GACACGCGCGCTCTGCTGAT/AGTGCAGAAGGGGCTGTGGG</td>
<td>510</td>
<td>ENSORLG00000009400 3</td>
<td>401 0</td>
</tr>
<tr>
<td>c-ski</td>
<td>CGAGACTGGGATGATCTTGTGTCG</td>
<td>491</td>
<td>ENSORLG00000016855 7</td>
<td>408 0</td>
</tr>
<tr>
<td>kita</td>
<td>CATGACTCTGTCCTGGGMAT/GCTAGAAGACCTCATTGCTTTCG</td>
<td>611</td>
<td>ENSORLG00000000569 4</td>
<td>237 270</td>
</tr>
<tr>
<td>mitta</td>
<td>CCTGCAATAGGACGARCTGAGGATGGG</td>
<td>456</td>
<td>ENSORLG0000003123 5</td>
<td>25 373</td>
</tr>
<tr>
<td>Tyr</td>
<td>TGGGTTGAGCGAATCTTCTCCCT/TGCGAATCCTGGCTATGTTG</td>
<td>659</td>
<td>ENSORLG00000010905 13</td>
<td>155 413</td>
</tr>
<tr>
<td>csfr1</td>
<td>AAGGACCGCTCTGGGAACGCACTG/GTACGACTGCTGTGCTGCCCTT</td>
<td>459</td>
<td>ENSORLG00000004849 10</td>
<td>25 324</td>
</tr>
<tr>
<td>pax3</td>
<td>AAGAGCCGCTGAGAAGCCGACGGAA/GTACGACTGCTGTGCTGCCCTT</td>
<td>471</td>
<td>ENSORLG00000015932 17</td>
<td>254 130</td>
</tr>
<tr>
<td>hag</td>
<td>AAATCGTGACTGCTCCCTGAGCT/AGGACGAGCAGCTACCTTGT</td>
<td>470</td>
<td>ENSORLG0000000906 15</td>
<td>115 309</td>
</tr>
<tr>
<td>rag</td>
<td>TCGCAGCTCTCCCTTGATGCT/TTGCGCCTGGGAAGGGG</td>
<td>461</td>
<td>ENSORLG0000011969 6</td>
<td>373 0</td>
</tr>
<tr>
<td>b2m</td>
<td>GCACTGACGATAGATTCCACCACCAGCT/ACGCTAYACRGYGGACYCTGTA</td>
<td>508</td>
<td>ENSORLG0000012506 23</td>
<td>235 183</td>
</tr>
<tr>
<td>gapdh</td>
<td>CCTGCGCAGATCCACCCCA/GACGTAYACRGYGGACYCTGTA</td>
<td>499</td>
<td>ENSORLG0000006033 16</td>
<td>171 258</td>
</tr>
<tr>
<td>cng1</td>
<td>CTGCTCTCCTGAGCGCTCCTGATTGC/ACGTGACTGCTGTGCTGCCCTT</td>
<td>707</td>
<td>ENSORLG0000005817 10</td>
<td>210 444</td>
</tr>
<tr>
<td>prl-like</td>
<td>CCCTGGGTGATGTCGTCGGGC/ACCCAGACAGGAGCCTTCCA</td>
<td>436</td>
<td>ENSORLG0000015652 24</td>
<td>368 0</td>
</tr>
<tr>
<td>enc1</td>
<td>CCRGCTTCCTGCGCCTTGGG/TCGGTGACCGGATGATCAG</td>
<td>417</td>
<td>ENSORLG00000035288 12</td>
<td>329 0</td>
</tr>
<tr>
<td>s7</td>
<td>CTGGCAATTTAATTCTGGACTG/AAACTCGTGYGGCTTTCTG</td>
<td>569</td>
<td>ENSORLG00000018123 24</td>
<td>0 414</td>
</tr>
</tbody>
</table>

attribute in part to the existence of alleles with different lengths resulting in double peaks (note: this is not an issue when using next-generation sequencing techniques). The percentage of missing data and gaps per species are listed in Table 2.

In total, we obtained a concatenated data set of 24 partial gene sequences containing 9669 bp. A total of 583 sites were variable (6.03%), of which 130 are parsimony-informative sites (1.3%) (calculated without the outgroup taxon *Tylochromis polyepis* and without indels or polymorphic sites coded with ambiguous IUPAC code). The combined sequence matrix consists of 5761 bp (59.58%) from exons and 3908 bp (40.42%) from introns (Table 1).

In 18 of the 24 loci, we detected heterozygous SNPs (46 SNPs in total; referred to as polymorphic sites in the tables, Table 2 and Table S1, Supporting information). Detailed information about the number of variable sites, the number and location of polymorphisms, the calculated BIC value and the gene ontology (GO) terms for every gene are listed in Table S1 (Supporting information). The PCR conditions for each primer pair (including the enzymes used and the annealing temperatures), and PCR and sequencing success (and possible reason for its failure) are shown in Table S2 (Supporting information).

Maximum likelihood (not shown) and Bayesian inference (Fig. 1) of the concatenated data yielded congruent trees. The only differences between the two trees concern weakly supported nodes (the relative positions of the Cyprichromini, Cyphotilapiini, Limnochromini, Eretmodini and Perissodini to the Lamprologini/Ectodini complex) and the placement of *Sargochromis* within the Haplochromini.

**Discussion**

In this study, we present a new primer set for phylogenetic inferences in East African cichlid fishes. We further show that our primers amplify successfully in most of the tested representatives from Lake Tanganyika, making our primer set applicable for a great portion of the ~250 cichlid species in this lake. As the primers amplify very well in our riverine and Tanganyikan representatives of haplochromines *sensu lato* (Salzburger et al. 2005), it is likely that they also work for the members of
the species flocks of Lake Malawi and the Lake Victoria region, which exclusively consist of haplochromines (Meyer et al. 1990; Verheyen et al. 2003; Salzburger et al. 2005).

We designed the primer pairs specifically for the use in both single-read Sanger sequencing and next-generation sequencing with the 454 technology, by restricting the amplicon product length to the read length of these methods (see Table 1). Another important attribute of our primers constitutes the similar melting temperature and consequently a comparable annealing temperature, which enables multiplexed reactions and cost-effective parallel high-throughput sequencing. Furthermore, the markers are placed into annotated genes with known functions, leading to a good coverage of the genome (see Table 1) and the avoidance (or at least relatively easy identification) of paralogs or pseudogenes. Finally, the primers were designed to cover coding (exons) and non-coding (introns) regions of these genes. Introns sequences have been used successfully for both phylogenetic inference (Hedin & Maddison 2001; Fujita et al. 2004; Dalebout et al. 2008; Jacobsen & Omland 2011; Yu et al. 2011) and population genetics (Palumbi & Baker 1994; Tay et al. 2008; Carvajal-Vallejos et al. 2010), particularly because

### Table 2: DNA sequencing success in our test taxon set.

<table>
<thead>
<tr>
<th>Taxon</th>
<th>Sequence length (bp)</th>
<th>Missing (%)</th>
<th>Gaps (%)</th>
<th>SNPs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bathybates graueri</td>
<td>9528</td>
<td>37.2</td>
<td>1.5</td>
<td>0</td>
</tr>
<tr>
<td>Cyprichromis leptosoma</td>
<td>9616</td>
<td>54.7</td>
<td>0.5</td>
<td>1</td>
</tr>
<tr>
<td>Ophthalotilapia ventralis</td>
<td>9517</td>
<td>56.9</td>
<td>1.6</td>
<td>7</td>
</tr>
<tr>
<td>Oreochromis tanganicae</td>
<td>9552</td>
<td>31.9</td>
<td>1.2</td>
<td>1</td>
</tr>
<tr>
<td>Cyphotilapia frontosa</td>
<td>9532</td>
<td>52.2</td>
<td>1.4</td>
<td>1</td>
</tr>
<tr>
<td>Eretmodus cyanostictus</td>
<td>9437</td>
<td>14.4</td>
<td>2.4</td>
<td>5</td>
</tr>
<tr>
<td>Tropheus moori</td>
<td>9482</td>
<td>7.3</td>
<td>1.9</td>
<td>4</td>
</tr>
<tr>
<td>Astatotilapia burtoni</td>
<td>9480</td>
<td>1.9</td>
<td>1.9</td>
<td>2</td>
</tr>
<tr>
<td>Lamprologus callipterus</td>
<td>9593</td>
<td>18.6</td>
<td>0.8</td>
<td>5</td>
</tr>
<tr>
<td>Neolamprologus pulcher</td>
<td>9530</td>
<td>31.8</td>
<td>1.4</td>
<td>1</td>
</tr>
<tr>
<td>Tylochromis polylepis</td>
<td>9633</td>
<td>37.9</td>
<td>0.4</td>
<td>2</td>
</tr>
<tr>
<td>Limnochromis abeelei</td>
<td>9494</td>
<td>13.1</td>
<td>1.8</td>
<td>5</td>
</tr>
<tr>
<td>Pseudocrenilabrus philander</td>
<td>9523</td>
<td>11.9</td>
<td>1.5</td>
<td>7</td>
</tr>
</tbody>
</table>

![Fig. 1](image-url)  
Bayesian inference topology inferred with MrBayes for the 16 species in our test data set and based on 24 genes (concatenated, 9669 bp). Bootstrap support for ML > 50 and posterior probability of the MrBayes analysis > 0.50 are shown. Branch lengths are proportional to the number of mutations per site.
introns typically contain a higher percentage of polymorphic sites than exons (reviewed in Zhang & Hewitt 2003). About 40% of the nuclear DNA sequences, obtained by using our newly developed primer set for cichlids, belong to introns, which indeed show greater diversity than the exons (see Table S1, Supporting information).

The primary goal of this novel primer set is to use it for phylogenetic purposes in order to refine and extend existing phylogenetic hypotheses (Salzburger et al. 2002b, 2005; Clabaut et al. 2005; Day et al. 2008; Koblmüller et al. 2008) and to address the gene tree/species tree issue in Tanganyikan cichlids (see Brito & Edwards 2009; Heled & Drummond 2010; Liu 2008 for methods and discussions).

The various drawbacks of the sole use of mitochondrial markers for phylogenetic, phylogeographic and population genetic inference have been frequently discussed (Ballard & Whitlock 2004; Ballard & Rand 2005; Rubinoff & Holland 2005; Brito & Edwards 2009; Galtier et al. 2009). However, also nuclear markers have some drawbacks, such as a relatively low mutation rate (Moritz et al. 1987) and a four times larger effective population size compared with the haploid and unparentally inherited mitochondrial markers leading to longer coalescence times and slower fixation rates (Moore 1995). In the case of the new marker set provided here, this is counterbalanced by the relatively large amount of sequence data that can be obtained.

Taken together, we present the development and the proof of functionality of the so far largest set of independent sequence-based nuclear markers for phylogenetic purposes for East African cichlid fishes. The markers can be used in both Sanger sequencing and next-generation sequencing using the 454 approach. We thus provide an important tool that will be used for multimarker phylogenetic analyses of East African cichlids in the future.

Acknowledgements

We are grateful to M. Barluenga, F. Muenzel, A. Indermaur and M. Muschick for help during fieldwork, B. Aeschbach and N. Boileau for technical assistance and A. Indermaur and M. Roesti for valuable comments on the manuscript. We thank three anonymous reviewers for helpful feedback on the manuscript. Sampling was performed under a memorandum of understanding (MOU) with the University of Zambia in Lusaka and the Department of Fisheries, Lake Tanganyika branch, Mpalungu, Zambia. This work was supported by the European Research Council (ERC; Starting Grant ‘INTERGENADAFT’), the Swiss National Science Foundation (grant 122438) and the University of Basel.

References


Koblmüller S, Duftner N, Sefc K et al. (2007a) Reticulate phylogeny of gastropod-shell-breeding cichlids from Lake Tanganyika – the result of repeated introgressive hybridization. *BMCEvolutionary Biology*, 7, 7.


Rambaut A (2009) FigTree v1.3.1. Available at: http://tree.bio.ed.ac.uk/software/figtree/.


B.S.M., W.S. conceived and designed the project. B.S.M. performed the experiments. B.S.M., W.S. analyzed the data and wrote the paper.

Data Accessibility

All DNA sequences from this study are available under GenBank Accession: JX135129–JX135389 (for more details see: Table S3, Supporting information).

Supporting Information

Additional supporting information may be found in the online version of this article.

Table S1 List of the 24 genes used for primer design. Information is provided with regard to the alignment length used for primer design; the percentage of variable and parsimony-informative sites without the outgroup; the number of polymorphic sites (SNPs), the location in the different species and its percentage including the outgroup; average p-distance calculated with complete deletion and without the outgroup; its range and SE calculated with 500 bootstrap replicates; and average p-distance including only base pairs from exon or intron without the outgroup; the recommended model from JMODELTEST 0.1.1 on the basis of the BIC with its computed likelihood scores; the GO terms from annotated fish sequences from UniProtKB.

Table S2 List of the 24 genes used for primer design. Information about the number of successful PCR and sequencing reactions; further used sequences and their Accession number of GenBank or other source; the percentage of missing data (due to sequencing errors); used annealing temperature in the PCR and used *Taq* polymerase; +/+ designates successful PCR and sequencing reaction, −/− both unsuccessful, +/−* PC*R successful and sequencing unsuccessful; indicating possible reason for failed sequencing reaction.

Table S3 List of used species, 24 markers and GenBank Accession numbers (asterisks indicate the usage of other unpublished primer pairs) http://www.ncbi.nlm.nih.gov/genbank/.

Please note: Wiley-Blackwell is not responsible for the content or functionality of any supporting information supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.