Comparative analysis of mosaic genomes among Lake Malawi cichlids

Yong-Hwee E. Loh[1], Lee S. Katz[1], Meryl C. Mims[1], Thomas D. Kocher[2], Soojin Yi[1] and J. Todd Streelman[*1]

[1] School of Biology, Petit Institute for Bioengineering and Bioscience, Georgia Institute of Technology, Atlanta, GA 30332, USA

[2]Department of Biology, University of Maryland, College Park, MD 20742, USA

[*]Corresponding author: J. Todd Streelman, Tel: (404) 385-4435; Fax: (404) 385-4440, Email: todd.streelman@biology.gatech.edu

Abstract

Background

Cichlid fishes from East Africa are remarkable for phenotypic and behavioral diversity on a backdrop of genomic similarity. Lake Malawi is home to the most species rich assemblage of African cichlids; as many as 800 – 1000 species are thought to have evolved from a common ancestor in the last 500K to 1MY. In 2006, the Joint Genome Institute completed low coverage survey sequencing of the genomes of five phenotypically and ecologically diverse Lake Malawi species. Here we report a computational and comparative analysis of these data.

Results

We produced assemblies for the 5 species ranging in aggregate length from 71 – 83 Mb, identified putative orthologs for over 12,000 human genes, and predicted more than 32,000 cross-species single nucleotide polymorphisms (SNPs), with ~2700 located in genic regions. Nucleotide diversity (Watterson's $\theta_w = 0.26\%$) was lower than that found among laboratory strains of the zebrafish. Jukes-Cantor genetic distance between species ranged from 0.23 – 0.29%, about one fifth of that between human and chimp. We collected ~18,000 genotypes to validate a subset of SNPs within and among populations and across multiple individuals of ~75 Lake Malawi species, and demonstrate the general utility of these markers.

Conclusion

Lake Malawi cichlids are mosaics of ancestrally polymorphic genomes. This assemblage of species presents a case of complex and dynamic evolutionary diversification, where recombination and the sorting of ancestral polymorphism may be more important than new mutation as sources of genetic variation. The unique mosaic structure of Lake Malawl cichlid genomes should facilitate conceptually new experiments, employing SNPs to identity genotype-phenotype association, using the entire species flock as a mapping panel.

Background

Cichlid fishes from the East African Rift lakes Victoria, Tanganyika and Malawi represent a preeminent example of replicated and rapid evolutionary radiation [1]. This group of fishes is a significant model of the evolutionary process and the coding of genotype to phenotype, largely because tremendous phenotypic diversity has evolved in a short period of time among lineages with similar genomes [2,3,4]. Recently evolved cichlid species segregate ancestral polymorphism [5,6] and may exchange genes [7,8]. Recent reports have capitalized on the diversity among East African cichlids to study the evolution and genetic basis of many traits, including behavior [9], olfaction [10], pigmentation [11,12,13], vision [14,15], sex determination [13,16], the brain [17] and craniofacial development [18,19,20].

Numerous genomic resources have been developed for East African cichlids (many of which are summarized at www.cichlidgenome.org). These include: genetic linkage maps for tilapia [13,21,22] and Lake Malawi species [18,20]; fingerprinted bacterial artificial chromosome libraries [23]; EST sequences for Lake Tanganyika and Lake Victoria cichlids (http://compbio.dfci.harvard.edu/tgi/tgipage.html); and first-generation micro-arrays [24,25]. Hundreds of studies have used these resources to study cichlid population genetics, molecular ecology, and phylogeny (reviewed in 26,27).

In 2006, under the auspices of the Community Sequencing Program, the Joint Genome Institute completed low coverage survey sequencing of the genomes of five Lake Malawi species. Species were chosen to maximize the morphological, behavioral and genetic diversity among the Malawi species flock. Here, we report computational and comparative analyses of these sequence data. We had three major goals: (i) to produce a low coverage assembly for each of the 5 species, (ii) to identify orthologs of vertebrate genes in these data and (iii) to predict single nucleotide polymorphisms (SNPs) segregating between species. Consequently, we produced assemblies for the 5 species ranging in aggregate length from 71 – 83 Mb, identified putative orthologs for over 12,000 human genes, and predicted more than 32,000 cross-species segregating sites (with ~2700 located in genic regions). Furthermore, we genotyped a test set of these SNPs in numerous Lake Malawi cichlid species and demonstrate the broad utility and resolution of these markers. Our work should facilitate further understanding of evolutionary processes in the species flocks of East African cichlids.

Results

Sequence assembly

Trace sequences of five Lake Malawi cichlid species, *Copadichromis conophorus* (CC), *Labeotropheus fuelleborni* (LF), *Melanochromis auratus* (MA), *Metriaclima zebra* (MZ) and *Rhamphochromis esox* (RE), were downloaded from the GenBank Trace Archive and assembled into contiguous (contig) sequences. The average cichlid genome is 1.1×10^9 bases [28] so the traces represent a sequence coverage of 0.12X to 0.17X for each of the 5 species (Supplementary Table 1). Through several quality filtering and assembly steps (Methods), the resultant genomic assemblies of the five cichlid species yielded an average of 61,533 contigs with a mean length of 1230 bases per contig. The total first-pass assembly sequence length for each species ranged from 71,315,231 bases (MA) to 83,266,025 bases (MZ), or about 7% of an average cichlid genome. Assembly statistics are shown in Table 1.

We noted that these first-pass assemblies were ‘over-assembled’ by roughly a factor of 2 when compared to theoretical expectations [29]. Theory suggests that random shotgun sequencing of single copy DNA, at 0.15X coverage of a 1.1 Gb genome, will result in an assembly length of ~153 Mb. We reasoned that our assemblies might be shorter than expected because multi-copy elements were grouped as if they were single copy sequence. Given the theoretical expectation (again for 0.15X coverage of a 1.1 Gb genome) that individual bases should only be sequenced a maximum of 4-5 times, we examined whether contigs were built from 5 or more trace sequences contributing overlapping bases. We observed that ~10 Mb of each first-pass assembly were derived from such contigs, and excluded these data from subsequent analyses (e.g., SNP prediction, see below). Notably, individual sequences contributing to these ‘high trace number’ contigs were not identified by RepeatMasker but did sometimes have Blast matches to putative repetitive elements (e.g., pol polyprotein, reverse transcriptase). Because of the keen interest in repetitive DNA families in cichlids [30] and other organisms [31], we have retained alignments of these ‘high trace number’ contigs and have marked them as such (see Supplementary Tables 3, 4).

Gene content and coverage

To establish the extent of gene content and coverage present in each assembly, we carried out BLASTX similarity searches (10^{-10} E-value cutoff) for each of the 5 assemblies against a reference human proteome (RefSeq proteins). The average proportion of putative genic sequence amounted to 3.7% of the available genomes. The MZ assembly contained the highest gene coverage, possessing genic loci that were significantly similar to approximately 5,240 unique human proteins. The remaining four species yielded approximately similar numbers ranging from 5,020 to 5,170 genes. It must be noted however that most of these genes are highly fragmented and incomplete, due to the low coverage of the assembly. In all, a total of 36% (12,211 genes out of 34,180; Supplementary Table 2) of the reference human proteome could be identified in one or more of the cichlid species.

Clustering and alignment

We obtained 25,458 clusters of putatively orthologous sequences, which were individually assembled into multi-species alignments for subsequent comparative analyses. Genic regions, as identified by similarity searches to known human and fish genes, were marked onto each alignment. Figure 1 illustrates a typical example of one such alignment.

Roughly 1% of the alignments (294 alignments) showed percentages of variable sites above 2% (~10 fold higher than the average). It is impossible to know, given the low coverage of the sequenced genomes, whether these represent orthologous but divergent regions of cichlid genomes or the alignment of paralogous sequence. We therefore retained these alignments, and included a calculation of polymorphism for each alignment (Supplementary table 3), for the consideration of researchers using these data. For example, alignment 108866 contains sequence with similarity to asteroid homologue 1, with 8% of sites variable and a majority of replacement polymorphism. Given the lack of functional information about this novel signaling protein (first described in *Drosophila*, ref. 32), this alignment provides useful information even if (and perhaps because) it includes paralogous loci. Another 12% of the alignments (2,119 total) contained individual species contigs that had consensus base positions derived from five or more trace sequences (see above).

For all subsequent analyses, we excluded 2,413 alignments that exhibited (i) a high percentage of variable sites and/or (ii) higher than expected coverage.

More than 11.6 million bases of multiple species alignments remain, of which roughly 1.06 Mb were inferred as genic. This included 10,902,011 (986,506 genic) bases of 2-species alignments, 721,049 (75,371 genic) bases of 3-species alignments, 27,951 (2,898 genic) bases of 4-species alignments and 877 (193 genic) bases of alignments containing all five species.

Segregating sites

Further analysis of these 11.6 million bases of multiple alignments identified a total of 32,417 (0.28%) cross-species single nucleotide polymorphisms (SNPs). In order to classify the quality of an identified variable site, a polymorphism quality score (PQS) was defined, corresponding to the first digit of the lowest Phrap quality score among the nucleotides of the different species present at the polymorphic site (e.g., a polymorphic site between 4 species with base quality scores of 34, 45, 46 and 50 would be assigned a PQS of three). In total, 4,468 (13.8%) variable sites had a PQS of five or higher, 7,952 (24.5%) had a PQS of four, 8,236 (25.4%) a PQS of three, and the remaining 11,761 (36.3%) had a PQS of two. PQS for each variable site are provided on the alignments described in Supplementary Table 3 (also, http://cichlids.biology.gatech.edu). Nucleotide diversity (Watterson's θ_W) averaged over 2-, 3- and 4-species alignments was 0.00257. Roughly 8% of all polymorphic sites (2,709) were located within the putative genic regions identified earlier. Alignments with fish and human proteins provided us with the phase information required to further classify these into 1,066 synonymous and 1,643 non-synonymous SNPs. Summaries of all alignments containing genic and non-genic polymorphisms are provided in Supplementary Tables 3 and 4.

In order to investigate the pairwise differences between any two species, all sequence alignment segments with 2 or more species were broken up into all possible pairwise alignments; this resulted in 1.06 – 1.55 Mb of alignment per pair. We then calculated the Jukes-Cantor distance between species pairs. The three shortest distances were between LF and MZ (0.229%), followed by MA/MZ (0.232%) and LF/MA (0.241%). A neighbor-joining tree of pairwise J-C distances strongly supports a clade made up of rock-dwelling (mbuna) species MA, MZ and LF (Figure 2), a result consistent with published analysis [3,4,33], but provides little resolution among mbuna genera (also consistent with previous work, ref. 12,34). We also calculated the ratio of replacement to synonymous substitutions (K_a/K_s) for concatenated genic alignments among all pairs of species. K_a/K_s ranged from 0.380 in CC/LF to 0.562 in LF/MA.

Validation and generality of SNPs

We genotyped 38 of our predicted SNPs, along with positive controls, in 384 Lake Malawi cichlid samples, using Beckman Coulter SNPstream™ technology. Positive controls were genes sequenced by others, with known variation in Malawi cichlids: *mitf*, *ednrb*, *aim1* and opsins *rh1*, *sws1*, *lws*, *sws2a* [3,14]. Predicted SNPs were chosen in this experiment if they showed sequence similarity to regions of *Tetraodon* (pufferfish) chromosome 11; we have previously shown *Tetraodon* 11 to share orthologs with cichlid chromosome 5

[20]. Our validation strategy sought to document the general use and segregation of these markers among Lake Malawi cichlids. Given recent divergence times among species (some as recent as 1000 years, ref. 2), we expected that SNPs might segregate throughout the assemblage. Our Malawi samples comprised ~10 individuals from each of 10 populations of MZ and LF, as well as 1-5 individuals of 77 additional species (25 of which were mbuna). Taxa were included to represent the morphological, functional and behavioral diversity of the Malawi lineage, which may contain more than 800 species [35].

Eight of 38 predicted polymorphisms were fixed (i.e., no variation) in all samples, indicating an error in sequencing (or genotyping), an error in prediction or the presence of a low frequency allele in the sequenced samples. Five predicted SNPs did not produce data reliable enough for genotype calls. The remaining 25 loci from our SNP predictions (66%) were polymorphic across the data set (Table 2). When taken together, these loci (plus positive controls) support previously reported population structure in MZ [36,37] and LF [38], as well as the genetic distinction between these species (MC Mims, unpublished). We mapped 5 predicted SNPs (*csrp1*, *sws2b*, *sema3f*, snp33, snp39) in the F_2 generation of an intercross between LF and MZ (JT Streelman, unpublished) and demonstrated Mendelian inheritance. As expected, these markers mapped to cichlid chromosome 5, which contains regions homologous to *Tetraodon* chromosome 11.

These SNP primer-probe combinations amplified DNA and detected biallelic polymorphism in over 75 diverse Lake Malawi cichlid species. Strikingly, LF and

MZ were never alternately fixed for SNP alleles, nor were mbuna-specific alleles present at any of the loci scored (Table 2). Malawi cichlid genomes are therefore highly similar (Jukes-Cantor distances of < 0.3%), and also segregate alleles widely throughout the flock. To visualize this further, we utilized a Bayesian approach that assigns individuals to a predefined number of genetic clusters [39]. Specifically, we were interested in how species would be assigned to major Malawi cichlid lineages identified in previous studies [3,4,33]. There are three such groups supported by the majority of molecular data: (i) the rock-dwelling mbuna, (ii) pelagic and sand-dwelling species, and (iii) a group comprised of *Rhamphochromis*, *Diplotaxodon* and other deep-water taxa. Analysis of the 31 SNP loci from Table 2 accurately classifies species to respective lineages (Figure 3). For instance, all species presently considered mbuna (blue) cluster with other mbuna, largely to the exclusion of other groups. Notably, the genomes of some species appear to be mosaics assembled from combinations of two, or even all three, major lineages. A few are worth comment. *Labidochromis gigas* (mbuna) has contributions from all three major lineages. All *Melanochromis* species investigated (*vermivorus*, *auratus*, *parallelus*) have roughly equal proportion mbuna and non-mbuna (green) genomes. Sand- or intermediate habitat-dwelling mbuna species (*Metriaclima livingstoni* and *M. patricki*, *Pseudotropheus crabro*) are combinations of mbuna and non-mbuna genomes. Some *Nimbochromis*, *Taeniolethrinops, Maravichromis* (*Mylochromis*), *Protomelas* and *Copadichromis* species are represented by individuals of both or varying proportions of non-mbuna genomes (red and green). Finally, species thought to represent the earliest divergence within the species flock (*Rhamphochromis* and the non-endemic *Astatotilapia calliptera*) carry contributions from mbuna and the more common non-mbuna genomes (red).

Discussion

African cichlid fishes are important models of evolutionary diversification in form and function [37]. They are singularly remarkable for the extent of phenotypic and behavioral diversity on a backdrop of genomic similarity. Lake Malawi is home to the most species rich assemblage of African cichlids; as many as 800 – 1000 species are thought to have evolved from a common ancestor in the last 500K to 1MY [35]. These recently formed species segregate ancestral polymorphism and exchange genes by hybridization [5,7,40]. Such circumstances present both opportunities and challenges for understanding evolutionary history and biological diversity. Opportunistically, researchers have used molecular markers across studies to interrogate the genetic basis of phenotypic differentiation [11,13,19,20]. This approach views Malawi cichlid species as natural mutants screened for function by natural selection; with essentially identical ancestral genomes honed by contrasting historical processes. By contrast, the task of reconstructing a phylogeny of species has been hindered by the very same phenomena of genomic similarity and mosaicism [2,3]; even the promising approach of AFLP does not provide strong resolution of the relationships among genera [12,34,41]. The data we present here should provide new resources and perspectives for cichlid evolutionary genomics.

Cichlid species are genomic mosaics

Lake Malawi cichlid species sequenced by the JGI embody the phylogenetic, morphological and behavioral diversity found within the assemblage. *Rhamphochromis esox* is a large (~0.5m) pelagic predator representing one of the basal lineages of the species flock [3,4,33]. *Copadichromis conophorus* is a sand-dwelling species that breeds on leks where males construct 'bowers' to attract females. *Melanochromis auratus*, *Metriaclima zebra* and *Labeotropheus fuelleborni* are rock-dwelling (mbuna) species that differ in color pattern, trophic ecology, body shape and craniofacial morphology (for pictures, see http://malawicichlids.com/index.htm).

Our data confirm the conclusions from previous genetic analyses on a smaller scale: Lake Malawi species have similar genomes. Jukes-Cantor genomic distances range from 0.23 – 0.29%, or roughly one fifth of this measure between human and chimpanzee (calculated from neutral sequence, ref. 42). Remarkably, the nucleotide diversity observed among the 5 cichlid species (Watterson's θ_w = 0.26%) is less than that found among laboratory strains of the zebrafish, *Danio rerio* (Watterson's θ_w = 0.48%, ref. 43). Although overall nucleotide diversity is less than that observed in *Danio*, the ratio of replacement to silent change is nearly 5-fold higher in the Lake Malawi genomes. Such a result might suggest that East African cichlid evolution is characterized by adaptive molecular evolution, as has been indicated in a few instances [14,44], or a relaxation of purifying selection attributable to small effective population size. However, we should view this estimate of K_a/K_s with caution, because of one of the remarkable features of these data (below). Variable sites identified from cross-species alignments are not substitutions fixed between species (Table 2, LF *vs*. MZ). The K_a/K_s approach to identifying selection may be largely inappropriate for such young species where ancestral alleles segregate as polymorphisms.

Despite phenotypic differences among the mbuna species MZ, MA and LF, the relationships among them remain difficult to parse (Figure 2). The pattern of variation observed across the ~75 species genotyped in this study demonstrates that biallelic polymorphisms segregate widely throughout the Malawi species flock (Table 2). SNPs segregate within and between MZ and LF populations, as well as within and among mbuna species and other lineages. No SNP locus surveyed is alternately fixed in LF versus MZ. In certain cases (e.g., *rhodopsin*, snp36), alleles are nearly mbuna-specific, but are observed in *Rhamphochromis* species. Lake Malawi cichlid species are mosaics of ancestrally polymorphic genomes (Figure 3). Add to this a propensity of recently diverged species to exchange genes [2], and Malawi cichlids present a case of complex and dynamic evolutionary diversification, where recombination and the sorting of ancestral polymorphism may be more important than new mutation as sources of genetic variation. Given the expectation of allele sharing across these young populations and species, the pattern of segregation for loci like *csrp1* (Table 2, LF *vs*. MZ) is worthy of further study.

Discovery for evolutionary biology

There are obvious challenges when attempting to extract information from low coverage genomic sequence, and also obvious payoffs [45,46,47]. Most previous studies have used this information for species-specific discovery (e.g., dog breeds) or broad evolutionary comparisons (e.g., dog-human, shark-human, cat-mammal). Our goals in the present analysis stem from the unique characteristics of Lake Malawi cichlids; these are biological species that behave genetically like a single population. Therefore, our biggest challenge was to devise a strategy that retains information from these low coverage survey sequences (0.75X spread over 5 closely related species), but minimizes error and bias in assembly and cross-species alignment for SNP identification. For example, we excluded many contigs because they appeared to be over-assembled, and we excluded multi-species alignments if they exceeded a polymorphism threshold. The over-assembly problem limits the coverage of these genomes in relation to expectation; this phenomenon, observed in the cat genome and in simulation, has complex and varying causes and has yet to be fully resolved [48]. It is likely to be mitigated to some degree by comparison to a higher-coverage reference sequence. The power of the data we present comes from the broad utility of the genic sequences and SNPs we have identified for many questions in genomic evolutionary biology.

Our analyses identified ~12,000 Lake Malawi cichlid sequences with similarity to human and fish proteins. This is a significant advance in our understanding of cichlid genomic content. To put this in context, approximately 13,500 unique ESTs, from 3 different East African cichlids, represent the sum total of such publicly released sequences (http://compbio.dfci.harvard.edu/tgi/tgipage.html). Our contribution roughly doubles the available data.

The ~32,000 (2,700 genic) SNPs we identified should provide a wealth of molecular markers for studies of population genetics and molecular ecology, linkage and QTL mapping, association mapping and phylogeny. We have shown these biallelic markers to be of general use, many segregating across the major cichlid lineages of Lake Malawi. Because SNP markers are (i) co-dominant, (2) easy to genotype, (3) reliable and reproducible from lab to lab and (4) readily mapped *in silico* (NHGRI will sequence a related cichlid, the tilapia, to 6X draft assembly coverage in 2008) they are likely to complement microsatellites and AFLP for most applications in cichlid evolutionary genomics. Given the unique mosaic structure of Lake Malawl cichlid genomes, it is exciting to envision experiments employing SNPs to identity genotype-phenotype associations, using the entire species flock as a mapping panel.

Methods

Samples

Individuals of *Copadichromis conophorus* (CC), *Labeotropheus fuelleborni* (LF), *Melanochromis auratus* (MA), *Metriaclima zebra* (MZ) and *Rhamphochromis esox* (RE), were sampled from the wild during an expedition to Malawi in 2005. Specimens prepared for survey sequencing by the JGI were collected from Mazinzi Reef (MZ), Domwe Island (LF, MA) and Otter Point (CC, RE), all locales in the southeastern portion of the lake. High-quality DNA was extracted and prepared in the laboratory of TDK.

Trace sequences

Trace sequences generated by the Joint Genome Institute (JGI) for CC, LF, MA, MZ and RE, together with their sequence quality scores, were downloaded (6 May 2007) from the NCBI Trace Archive. The dataset for each species consisted of an average of about 152,000 individual trace reads with total read lengths ranging from 137 – 185 million bases. Detailed sequence statistics for each species are provided in Supplementary Table 1.

Sequence pre-processing and assembly

The trace and quality sequences were first pre-processed for assembly by masking out all possible vector sequences available from the NCBI UniVec vector sequence database (downloaded 6 May 2007). The vector masking was performed using the cross_match.pl perl script provided by the Phred-Phrap package [49]. In order to reduce the computational complexity and time required for the final assembly, repeat sequences were masked prior to assembly using RepeatMasker version 3.1.8 (Smit, Hubley and Green, unpublished data) in conjunction with the latest repeatmasker libraries from RepBase Update [50]. Bases with sequencing quality score of less than 20 were also masked. The actual assembly of each species' trace sequences into contiguous sequences (contigs) was then performed using the Phrap version 0.990329 assembly program from the Phred-Phrap package. Contigs with more than 80% low quality bases (defined as <20 assembly quality score) were removed from the assembly. The genomic assembly sequences for each species have been deposited in GenBank under accession numbers XXXX.

Similarity search and alignment

Orthologous genomic contig pairs were first identified using reciprocal BLASTN similarity searches with a strict E-value cutoff of 10^{-100}, performed across the sequence contigs of all possible species pairs. To reduce spurious ortholog assignments, putative ortholog contig pairs were only retained if their regions of high sequence similarity (1) formed good end-to-end overlaps (defined as within 100 bases of the 5' end or 30 bases from the 3' end of a sequence), or (2) overlap more than 80% of the shorter contig. Though some of the filtered regions could represent biologically relevant loci where recombination or translocations might have occurred, we decided to remove them from this analysis. Contig pair assignments were then passed to an algorithm that created clusters of contigs whereby each contig within the cluster must be related to all other contigs in the cluster through one or more putatively orthologous relations. Each cluster of contigs was then individually aligned using Phrap, resulting in a continuous alignment tiling path where each alignment position may consist of a base from any one or up to all five cichlid species (Figure 1). Segregating sites were then identified from alignment positions with high quality bases (>20 score) from two or more species. A polymorphism quality score (PQS) was defined, corresponding to the first digit of the lowest Phrap quality score among the nucleotides of the different species present at the polymorphic site (e.g., a polymorphic site between 4 species with base quality scores of 34, 45, 46 and 50 would be assigned a PQS of three). To compare the extent of nucleotide diversity among the five cichlid species, we calculated Watterson's theta (θ_w, ref. 51). This measure takes into account the number of variable positions and the sample size analyzed. Our data violate the assumption of an infinite, interbreeding population, but we chose this metric to in order to make direct comparisons to similar measures from study of other genomes (e.g., zebrafish).

Protein-coding sequence identification

Cichlid protein coding sequences were inferred based on similarity searches to known protein databases of fishes and humans. BLASTX searches with E-value cutoff of 10^{-10} were performed for the each cichlid genomic assembly as well as the overall consensus sequence of the cluster alignments, against a protein database made up of all GenBank *Actinopterygii* (ray-finned fishes) sequences (downloaded 02 June 2007; 163,471 entries) and all human RefSeq proteins (downloaded 25 June 2007; 34,180 sequences). The alignment with the highest scoring hit for each genomic locus was then used as a reference to determine the coding strand and phase of the protein-coding cichlid locus.

Evolutionary divergence

All cluster alignment segments with contributing bases from two or more species were split into pairwise alignments (each 2, 3, 4 or 5 species alignment position can be split into 1, 3, 6 or 10 pairwise alignments respectively). Pairwise alignments within each of the 10 possible species pair combinations (CC-LF, CC-MA, CC-MZ, CC-RE, LF-MA, LF-MZ, LF-RE, MA-MZ, MA-RE, MZ-RE) were then concatenated and the number of substitutions counted. Jukes-Cantor correction for multiple substitutions was applied to these direct distance measurements [52]. A neighbor-joining phylogenetic tree [53] was reconstructed using the Neighbor program of the PHYLIP package (version 3.67, ref. 54).

To evaluate statistical support for this phylogenetic topology, 100 bootstrap replicates of multiple sequence alignments were generated, with each replicate sampling (with replacement) the exact numbers of 2, 3, 4 and 5-species alignment positions as in the original data set. The replicate alignments were then broken into pairwise alignments and divergences calculated as described. Neighbor-joining trees for each replicate alignment were constructed as before and a consensus tree was determined using the Consense program in PHYLIP.

Pairwise alignments consisting of only genic sequences were obtained from multi-species cluster alignment segments in a manner similar to that described above. The DNAStatistics package of Bioperl (www.bioperl.org) was then used to calculate the K_a/K_s values of pairwise alignments.

Genotyping, validation and visualization of mosaic genomes

A subset of SNPs (by manual inspection of JGI traces or the automated procedure described here) was chosen based on BLAST matches to *Tetraodon* chromosome 11 which has significant homology to cichlid chromosome 5 [20]. We chose to validate a modest number of SNPs in a wide diversity of Lake Malawi samples, given the manifold interests of the Malawi research community. The GenomeLab SNPstream Genotyping System Software Suite v2.3 (Beckman Coulter, Inc., Fullerton, CA) was used for experimental setup, data uploading, image analysis, genotype calling and QC review, at Emory University's Center for Medical Genomics. In brief, marker panel data (i.e., multiplexed SNP panel designed by SNPstream's Primer Design Engine website [www.autoprimer.com]) were first uploaded to the SNPstream database using the PlateExplorer application software. Also uploaded was the Process Group Data containing all test sample information generated through a Laboratory Information Management System (Nautilus 2002, Thermo Fisher Scientific, Waltham, MA). An on-board CCD camera of the SNPstream Imager took two snapshot images of each well of the 384-well tag array, one under a blue excitation laser, the other under a green excitation laser. Image application software was used to analyze the captured images to detect spots, overlay an alignment grid, and determine spot intensity. The fluorescent pixel intensity data for each SNP under the two channels, representing the relative abundance of the two alleles, were uploaded to the database. The GetGenos application software was used to calculate and generate a Log(B+G) *vs.* B/(B+G) plot, where B and G were the pixel intensities under the blue and green channels, respectively, for each sample and each SNP. Next, automated genotype calling was accomplished using the QCReview application software based on a number of criteria (e.g., signal baseline, clustering pattern of the three genotypes, Hardy-Weinberg score). A genotype summary was generated using the Report application software. We mapped 5 SNPs (*csrp1*, *sws2b*, *sema3f*, snp33, snp39) in the F_2 generation of an intercross between LF and MZ using PCR-RFLP.

To visualize the mosaic structure of individual genomes, we used a Bayesian assignment method (STRUCTURE v.2.2, ref. 39). We chose to define the number of K genetic clusters in accordance with previous research showing ~three major evolutionary groups of Lake Malawi cichlids [3,4,5,33]. Note that we do not intend this to mean that 3 is the best supported estimate of K in these data; our rationale is rather to demonstrate how individual genomes are composites (or not) of the major evolutionary lineages found in the lake. Thus, we used the <u>admixture</u> model to estimate q, the proportion of each genome derived from each of K genetic clusters.

List of abbreviations

CC	*Copadichromis conophorus*
LF	*Labeotropheus fuelleborni*
MA	*Melanochromis auratus*
MZ	*Metriaclima zebra*
RE	*Rhamphochromis esox*
SNP	Single Nucleotide Polymorphism
PQS	Polymorphic Quality Score

Acknowledgements

We thank members of the Streelman lab and Karen Carleton for comments on previous drafts of this manuscript. The research is supported by grants from the NSF (IOB 0546423), NIH (R21 DE017182) and Alfred P. Sloan Foundation (BR-4499) to JTS. Drs. Karen Carleton and Federica DiPalma extracted high quality DNA from the 5 species of Malawi cichlid. Library construction and sequencing was performed by the Joint Genome Institute under the auspices of the US Department of Energy's Office of Science, Biological and Environmental Research Program and by the University of California, Lawrence Livermore National Laboratory under Contract No. W-7405-Eng-48, Lawrence Berkeley National Laboratory under Contract No. DE-AC03-76SF00098 and Los Alamos National Laboratory under Contract No. W-7405-ENG-36.

Figure Legends

Figure 1. Alignment of a typical cluster of orthologous sequences. A. Overall alignment of assembly contigs from 3 different cichlid species with alignment positions indicated. B. Expanded detail of nucleotide alignment. Filled pink block shows the expanded alignment corresponding to dotted red box in A. Filled blue block shows the alignment of corresponding species' traces that made up the assembly sequences. Lowercase nucleotides have base quality scores <20. Alignment positions shown after consensus sequence. Dots "." represent identity in alignment. Cap "^" represents segregating site. PQS shown below A-G SNP site.

Figure 2. Pairwise Jukes-Cantor distance matrix between cichlid species and the neighbor-joining phylogenetic tree generated. Numbers indicate bootstrap percentages from 100 permutations.

Figure 3. Lake Malawi cichlids exhibit mosaic genomes. We show the contribution to each individual genome (q, which ranges from 0 – 100%) from each of K = 3 predefined genetic clusters (blue, red, green), for data derived from SNPs in Table 2. Note that this method predefines the number, but not the identity of genetic clusters. Species names are written once; multiple individuals from species are grouped together (e.g., 4 individuals of *Pseudotropheus crabro*). Notably, all mbuna genera and species have the majority, or near majority of their genomes from one major lineage (blue) and all non-mbuna draw

their genomes from the other two major lineages (red, green). Various species contain individuals with mosaic genomes (e.g., *Labidochromis gigas*, *Melanochromis* sp., *Metriaclima livingstonii*, *M. patricki*, *Pseudotropheus crabro*, *Astatotilapia calliptera*). Certain genera undergoing revision (*Copadichromis*, *Taeniolethrinops*, *Maravichromis* [*Mylochromis*], *Protomelas*) as well as *Nimbochromis* segregate cryptic variation among populations and individuals (OP is Otter Point, TW is Thumbi West Island). Note that according to this analysis of nuclear SNPs, there is genetic structure among non-mbuna species, but there is not a group that corresponds to *Rhamphochromis* and *Diplotaxodon* (a feature of most mtDNA gene trees).

Tables

Table 1. First-pass genomic assembly statistics for five Lake Malawi cichlid species.

	CC	LF	MA	MZ	RE
Total number of contigs in assembly	63,106	58,579	63,750	66,359	55,870
Total length (bases)	77,214,778	73,456,225	71,315,231	83,266,025	73,391,924
Genome coverage[a] (%)	7.02	6.68	6.48	7.57	6.67
Shortest contig length (bases)	40	46	42	32	44
Longest contig length (bases)	19,632	17,437	21,601	15,371	21,351
Mean contig length (bases)	1,224	1,254	1,119	1,255	1,314
Q25 contig length (bases)	792	873	820	840	970
Q50 (median) contig length (bases)	1,005	1,103	991	1,193	1,155
Q75 contig length (bases)	1,433	1,388	1,151	1,444	1,444
Total genic length (bases)	2,863,110 (3.7%)	2,841,933 (3.9%)	2,761,941 (3.9%)	2,851,968 (3.4%)	2,797,548 (3.8%)

[a] using an average cichlid genome size of 1.1×10^9 bases.

Table 2. Major allele frequency for biallelic SNPs surveyed across Lake Malawi cichlid populations and species. The first seven loci represent positive controls as explained in the text. Two SNPs were predicted and genotyped in *sws2b*; genotypes were in perfect linkage so only one is shown here.

snp/pop	***aim1***	***mitf***	***ednrb***	***rhodopsin***	***sws1***	***sws2a***	***lws***	***sws2b***	**snp8**	**snp10**
MZ otter	1	1	0.83	0.45	1	0.9	0.5	1	0.75	1
MZ chiofu	0.85	1	0.71	0.1	NA	0.7	1	1	0.95	1
MZ eccles	0.75	1	1	0	1	0.55	0.25	1	0.95	1
MZ masinge	0.78	1	0.75	0.95	1	1	1	1	0.5	1
MZ makanjila	0.85	1	0.875	0	1	0.95	1	0.95	0.75	1
MZ west	0.85	1	1	0	1	0.95	1	0.95	0.75	1
MZ mazinzi	0.35	0.9	0.66	0.81	1	0.8	0.3	1	0.5	1
MB mazinzi	0.43	1	1	1	1	0.64	0.86	1	1	1
MZ zimbawe	0.5	1	0.688	1	0.43	0.75	0.1	1	0.9	1
MZ domwe	0.8	1	0.5	0.85	1	0.95	0.45	1	0.75	1
LF west	0.84	0.875	0.4	0.14	1	1	1	1	0.5	1
LF otter	0.7	1	NA	1	1	1	1	0.6	0.7	1
LF chinyamwezi	1	1	1	1	1	1	1	1	1	1
LF chinyamkwazi	1	1	1	1	1	1	1	1	1	1
LF eccles	1	1	0	1	1	0.95	1	1	1	1
LF chiofu	0.75	1	0	1	1	1	1	1	0.75	1
LF makanjila	0.9	0.9	0.71	0.11	1	1	1	0.95	0.7	1
LF zimbawe	0.45	1	0	1	1	1	1	1	0.25	1
LF domwe	0.75	1	0	1	1	1	1	1	0.25	1
LF mumbo	0.89	1	0	1	1	0.95	1	0.88	0.6	1
All MZ	**0.73**	**0.989**	**0.8**	**0.48**	**0.95**	**0.82**	**0.65**	**0.989**	**0.78**	**1**
All LF	**0.83**	**0.984**	**0.37**	**0.82**	**1**	**0.99**	**1**	**0.96**	**0.7**	**1**
Mbuna (25 sp.)	**0.7**	**1**	**0.63**	**0.49**	**0.86**	**0.73**	**0.92**	**0.85**	**0.62**	**1**
Others (52 sp.)	**0.992**	**0.73**	**0.042**	**0.08**	**0.71**	**0.044**	**0.792**	**0.812**	**0.984**	**0.836**

snp/pop	**snp11**	**snp13**	***csrp1***	**snp19**	**snp21**	**snp22**	**snp23**	**snp24**	**snp25**	**snp27**
MZ otter	0.85	1	1	0.65	0.15	1	1	1	1	1
MZ chiofu	0.45	0.95	1	0.8	1	1	1	0.75	1	1
MZ eccles	0.19	1	1	0.4	0.7	1	1	0.95	1	1
MZ masinge	0.89	0.95	0.95	0.56	0.83	0.94	1	0.78	0.81	1
MZ makanjila	0.2	1	0.89	0.7	0.4	0.55	0.9	0.81	0.9	1
MZ west	0.25	1	0.9	0.9	0.45	0.75	1	0.8	1	1
MZ mazinzi	0.5	0.875	1	1	0.4	1	1	0.75	1	1
MB mazinzi	0.36	1	1	0.71	0.75	1	1	0.67	1	1
MZ zimbawe	0.06	0.95	0.8	0.89	0.5	1	1	0.75	1	1
MZ domwe	0.45	1	0.9	0.95	0.5	0.9	1	0.85	1	1
LF west	0.375	0.954	0.05	0	0.96	0.54	1	0.96	0.92	0.95
LF otter	0.4	1	0.1	0.3	1	0	0.9	1	0.8	0.9
LF chinyamwezi	0.15	1	0	0	1	0	1	0.55	0.44	1
LF chinyamkwazi	0.3	1	0	0.1	0.7	0.1	0.35	0.75	1	1
LF eccles	1	1	0	0.5	1	0.55	0.85	1	1	1
LF chiofu	0.9	1	0	0	1	0	0.75	1	1	1
LF makanjila	0.61	0.833	0.11	0.05	0.96	0.5	1	1	0.9	1
LF zimbawe	0.25	1	0	0.85	1	0.3	0.3	0.75	0.95	0.3
LF domwe	0.45	0.85	0.2	0.4	1	0.1	0.65	0.95	0.85	0.95
LF mumbo	0.45	1	0	0.7	1	0	0.95	0.5	1	1
All MZ	**0.43**	**0.978**	**0.946**	**0.76**	**0.55**	**0.91**	**0.989**	**0.82**	**0.97**	**1**
All LF	**0.51**	**0.962**	**0.04**	**0.27**	**0.96**	**0.22**	**0.77**	**0.85**	**0.88**	**0.93**
Mbuna (25 sp.)	**0.55**	**0.9**	**0.69**	**0.62**	**0.91**	**0.92**	**0.94**	**0.43**	**0.94**	**0.992**
Others (52 sp.)	**0.912**	**1**	**0.988**	**0.55**	**0.992**	**1**	**1**	**0.21**	**0.98**	**1**

snp/pop	**snp29**	**snp30**	**MSP**	***sema 3c***	***sema 3f***	**snp33**	**snp36**	**snp37**	**snp39**	**snp40**	**snp44**
MZ otter	1	0.65	1	0.95	0.8	0.45	1	1	0.8	1	0.95
MZ chiofu	1	0.1	1	0.9	0.35	0.95	1	1	0.55	1	0.9
MZ eccles	1	0.8	1	1	0.35	0.75	1	1	0.89	1	1
MZ masinge	1	0.94	1	1	1	0.84	0.95	1	0.94	1	0.88
MZ makanjila	1	0.85	0.95	1	0.5	0.95	1	1	0.7	1	0.95
MZ west	1	0.8	1	0.95	0.6	1	1	1	0.7	1	1
MZ mazinzi	1	0.61	1	0.94	0.95	1	1	1	0.67	1	0.6
MB mazinzi	1	0.75	1	1	0.86	0.93	1	1	0.93	1	1
MZ zimbawe	1	0.65	1	1	0.5	0.75	1	1	0.72	1	0.85
MZ domwe	1	0.5	1	0.65	0.6	0.65	1	1	0.7	1	0.9
LF west	0.95	1	1	0.96	0	0.46	1	0.75	1	1	0.83
LF otter	1	0.7	0.95	1	0	0	1	0.6	1	1	1
LF chinyamwezi	1	1	1	1	0	0	1	0.7	1	1	1
LF chinyamkwazi	1	1	1	0.9	0.45	0	1	0.1	1	1	1
LF eccles	1	0.33	1	0.67	0.9	0.22	1	0.5	1	1	1
LF chiofu	1	1	1	0.83	0	0	1	0.4	1	1	0.95
LF makanjila	0.95	1	1	0.9	0.1	0.39	1	0.89	1	1	0.94
LF zimbawe	1	1	1	1	0	0.17	1	0.35	1	1	1
LF domwe	1	0.75	1	0.95	0.3	0	1	0.11	1	1	1
LF mumbo	1	1	1	1	0	0.2	1	0.25	1	1	1
All MZ	**1**	**0.66**	**0.99**	**0.94**	**0.62**	**0.83**	**0.99**	**1**	**0.75**	**1**	**0.9**
All LF	**0.989**	**0.89**	**0.99**	**0.9**	**0.26**	**0.15**	**1**	**0.49**	**1**	**1**	**0.97**
Mbuna (25 sp.)	**1**	**0.86**	**0.97**	**0.86**	**0.54**	**0.62**	**0.79**	**0.99**	**0.9**	**1**	**0.92**
Others (52 sp.)	**0.95**	**0.05**	**0.75**	**0.87**	**0.02**	**0.984**	**0.03**	**1**	**0.992**	**0.989**	**0.992**

Supplementary Files

Supplementary Table 1. Trace sequence statistics of five Lake Malawi cichlid species.

Supplementary Table 2. Human gene homologs present in the five cichlid species.

Supplementary Table 3. List of alignment and polymorphic sites.

Supplementary Table 4. List of alignments with BLAST hits to fish and humans.

References

1. Kocher TD: **Adaptive evolution and explosive speciation: the cichlid fish model.** *Nat. Rev. Genet*. 2004, **5:** 288-98.

2. Won Y-J, Sivasundar A, Wang Y, Hey J: **On the origin of Lake Malawi cichlid species.** *Proc. Natl. Acad. Sci. USA* 2005, **102:** 6581-6586.

3. Won Y-J, Wang Y, Sivasundar A, Raincrow J, Hey J: **Nuclear gene variation and molecular dating of the cichlid species flock of Lake Malawi.** *Mol. Biol. Evol.* 2006, **23:** 828-837.

4. Hulsey CD, Mims MC, Streelman JT: **Do constructional constraints influence cichlid craniofacial diversification?** *Proc. Biol. Sci.* 2007, **274:** 1867-1875.

5. Moran P, Kornfield I: **Retention of ancestral polymorphism in the Mbuna species flock of Lake Malawi.** *Mol. Biol. Evol.* 1993, **10:** 1015-1029.

6. Nagl S, Tichy H, Mayer WE, Takahata N, Klein J: **Persistence of neutral polymorphisms in Lake Victoria cichlid fish.** *Proc. Natl. Acad. Sci. USA* 1998, **24:** 14238-14243.

7. Smith PF, Konings A, Kornfield I: **Hybrid origin of a cichlid population in Lake Malawi: implications for genetic variation and species diversity.** *Mol. Ecol.* 2003, **12:** 2497-2504.

8. Seehausen O: **Hybridization and adaptive radiation.** *Trends Ecol. Evol.* 2004, **19:** 198-207.

9. Aubin-Horth N, Desjardins JK, Martei YM, Balshine S, Hofmann HA: **Masculinized dominant females in a cooperatively breeding species.** *Mol. Ecol.* 2007, **16:** 1349-1358.

10. Blais J, Rico C, van Oosterhout C, Cable J, Turner GF, Bernatchez L: **MHC adaptive divergence between closely related and sympatric African cichlids.** *PloS ONE* 2007, **2:** e734.

11. Streelman JT, Albertson RC, Kocher TD: **Genome mapping of the orange blotch colour pattern in cichlid fishes.** *Mol. Ecol.* 2003, **12:** 2465-2471.

12. Allender CJ, Seehausen O, Knight ME, Turner GF, Maclean N: **Divergent selection during speciation of Lake Malawi cichlid fishes inferred from parallel radiations in nuptual coloration.** *Proc. Natl. Acad. Sci. USA* 2003, **100:** 14074-14079.

13. Lee B-Y, Lee W-J, Streelman JT, Carleton KL, Howe AE, Hulata G, Slettan A, Stern JE, Terai Y, Kocher TD: **A second-generation genetic linkage map of tilapia (*Oreochromis* spp).** *Genetics* 2005, **170:** 237-244.

14. Spady TC, Seehausen O, Loew ER, Jordan RC, Kocher TD, Carleton KL: **Adaptive molecular evolution in the opsin genes of rapidly speciating cichlid fishes.** *Mol. Biol. Evol.* 2005, **22:** 1412-1422.

15. Parry JW, Carleton KL, Spady T, Carboo A, Hunt DM, Bowmaker JK: **Mix and match color vision: tuning spectral sensitivity by differential gene expression in Lake Malawi cichlids.** *Curr. Biol.* 2005, **15:** 1734-1739.

16. Lee B-Y, Hulata G, Kocher TD: **Two unlinked loci controlling the sex of blue tilapia (*Oreochromis aureus*).** *Heredity* 2004, **92:** 543-549.

17. Huber R, van Staaden MJ, Kaufman LS, Liem KF: **Microhabitat use, trophic patterns, and the evolution of brain structure in African cichlids.** *Brain Behav. Evol.* 1997, **50:** 167-182.

18. Albertson RC, Streelman JT, Kocher TD: **Directional selection has shaped the oral jaws of Lake Malawi cichlid fishes.** *Proc. Natl. Acad. Sci. USA* 2003, **100:** 5252-5257.

19. Albertson RC, Streelman JT, Kocher TD, Yelick PC: **Integration and evolution of the cichlid mandible: the molecular basis of alternative feeding strategies.** *Proc. Natl. Acad. Sci. USA* 2005, **102:** 16287-16292.

20. Streelman JT, Albertson RC: **Evolution of novelty in the cichlid dentition.** *J Exp. Zoolog. B Mol. Dev. Evol.* 2006, **306:** 216-226.

21. Kocher TD, Lee W-J, Sobolewska H, Penman D, McAndrew B: **A genetic linkage map of the cichlid fish, the tilapia (*Oreochromis niloticus*).** *Genetics* 1998, **148:** 1225-1232.

22. Carleton KL, Streelman JT, Lee B-Y, Garnhart N, Kidd MR, Kocher TD: **Rapid isolation of CA microsatellites from the cichlid genome.** *Anim. Genet.* 2002, **33:** 140-144.

23. Katigiri T, Kidd CE, Tomasino E, Davis JT, Wishon C, Stern JE, Carleton KL, Howe AE, Kocher TD**: A BAC-based physical map of the Nile tilapia genome.** *BMC Genomics* 2005, **6:** 89.

24. Kijimoto T, Watanabe M, Fujimura K, Nakazawa M, Murakami Y, Kuratani S, Kohara Y, Gojobori T, Okada N**: cimp, a novel astacin family metalloproteinase gene from East African cichlids, is differentially expressed between species during growth.** *Mol. Biol. Evol.* 2005, **22:** 1649-1660.

25. Renn SC, Aubin-Horth N, Hofmann HA: **Biologically meaningful expression profiling across species using heterologous hybridization to a cDNA microarray.** *BMC Genomics* 2004, **6:** 42.

26. Kornfield I, Smith PF: **African cichlid fishes: model systems for evolutionary biology.** *Ann. Rev. Ecol. Evol. Syst.* 2000, **31:** 163-196

27. Genner MJ, Turner GF: **The mbuna cichlids of Lake Malawi: a model for rapid speciation and adaptive radiation.** *Fish and Fisheries* 2005, **6:** 1-34.

28. Gregory TR, Nicol JA, Tamm H, Kullman K, Leitch IJ, Murray BG, Kapraun DF, Greilhuber J, Bennett MD: **Eukaryotic genome size database.** *Nucleic Acids Res.* 2007, **35:** D332-D338.

29. Lander ES, Waterman MS: **Genomic mapping by fingerprinting random clones: a mathematical analysis.** *Genomics* 1988, **2**: 231-239.

30. Takahashi K, Okada N: **Mosaic structure and retropositional dynamics during evolution of subfamilies of short interspersed elements in African cichlids.** *Mol. Biol. Evol*. 2002, **19**: 1303-1312.

31. Jordan IK, Rogozin IB, Glazko GV, Koonin EV: **Origin of a substantial fraction of human regulatory sequences from transposable elements.** *Trends Genet*. 2003, **19**: 68-72.

32. Kotarski MA, Leonard DA, Bennett SA, Bishop CP, Wahn SD, Sedore SA, Shrader M: **The *Drosophila* gene *asteroid* encodes a novel protein and displays dosage-sensitive interactions with *Star* and *Egfr*.** *Genome* 1998, **41:** 295-302.

33. Kocher TD, Conroy JA, McKaye KR, Stauffer JR, Lockwood SF: **Evolution of NADH dehydrogenase subunit 2 in East African cichlid fish.** *Mol. Phylogenet. Evol.* 1995, **4:** 420-432.

34. Albertson RC, Markert JA, Danley PD, Kocher TD: **Phylogeny of a rapidly evolving clade: the cichlid fishes of Lake Malawi, East Africa.** *Proc. Natl. Acad. Sci. USA* 1999, **96:** 5107-5110.

35. Turner GF, Seehausen O, Knight ME, Allender CF, Robinson RL: **How many species of cichlid fishes are there in African lakes?** *Mol. Ecol.* 2001, **10:** 793-806.

36. Danley PD, Markert JA, Arnegard ME, Kocher TD: **Divergence with gene flow in the rock-dwelling cichlids of Lake Malawi.** *Evolution* 2000, **54:** 1725-1737.

37. Streelman JT, Peichel CL, Parichy DM: **Developmental genetics of adaptation in fishes: the case for novelty.** *Ann. Rev. Ecol. Evol. Syst.* 2007, **38:** 655-681.

38. Arnegard ME, Markert JA, Danley PD, Stauffer JR, Ambali AJ, Kocher TD: **Population structure and colour variation of the cichlid fish *Labeotropheus fuelleborni* along a recently formed archipelago of rocky habitat patches in southern Lake Malawi.** *Proc. Biol. Sci.* 1999, **266:** 119-130.

39. Pritchard JK, Stephens M, Donnelly P: **Inference of population structure using multilocus genotype data.** *Genetics* 2000, **155:** 945-959.

40. Streelman JT, Gmyrek SL, Kidd MR, Kidd CE, Robinson RL, Hert E, Ambali AJ, Kocher TD: **Hybridization and contemporary evolution in an introduced cichlid fish from Lake Malawi National Park.** *Mol. Ecol.* 2004, **13:** 2471-2479.

41. Kidd MR, Kidd CE, Kocher TD: **Axes of differentiation in the bower-building cichlids of Lake Malawi.** *Mol. Ecol.* 2006, **15:** 459-478.

42. Chen F-C, Vallender EJ, Wang H, Tzeng C-S, Li W-H: **Genomic divergence between human and chimpanzee estimated from large-scale alignments of genomic sequences.** *J. Hered*. 2001, **92:** 481-489.

43. Guryev V, Koudijs MJ, Berezikov E, Johnson SL, Plasterk RHA, van Eeden FJM, Cuppen E: **Genetic variation in the zebrafish.** *Genome Res.* 2006, **16:** 491-497.

44. Terai Y, Morikawa N, Okada N: **The evolution of the pro-domain of bone morphogenetic protein 4 (Bmp4) in an explosively speciated lineage of East African cichlid fishes.** *Mol. Biol. Evol.* 2002, **19:** 1628-1632.

45. Kirkness EF *et al.*: **The dog genome: survey sequencing and comparative analysis.** *Science* 2003, **310:** 1898-1903.

46. Venkatesh B *et al.*: **Survey sequencing and comparative analysis of the elephant shark (*Callorhinchus milii*) genome.** *PloS Biology* 2007, **5**: e101. doi:10.1371

47. Pontius JU, Mullikin JC, Smith DR *et al.*: **Initial sequence and comparative analysis of the cat.** *Genome Res*. 2007, **17**: 1675-1689.

48. Green P: **2x genomes – does depth matter?** *Genome Res*. 2007, **17**: 1547-1549.

49. Ewing B, Hiller L, Wendl M, Green P: **Basecalling of automated sequencer traces using Phred. I. Accuracy assessment.** *Genome Res.* 1998, **8:** 175-185.

50. Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J: **Repbase Update, a database of eukaryotic repetitive elements.** *Cytogenet Genome Res.* 2005, **110:** 462-467.

51. Watterson GA: **On the number of segregating sites in genetic models without recombination.** *Theor Pop Biol.* 1975, **7:** 256-276.

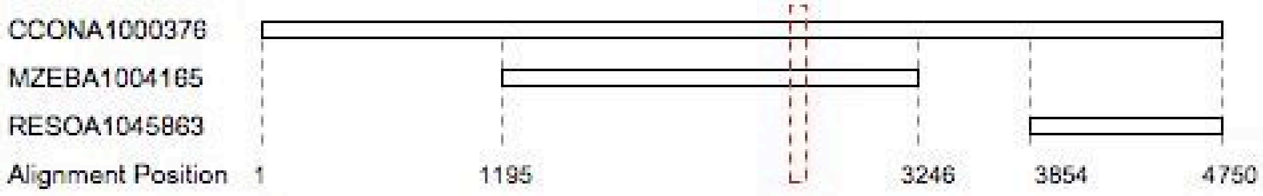
52. Jukes TH, Cantor CR: **Evolution of protein molecules**. In *Mammalian Protein Metabolism.* Edited by Munro HN. New York: Academic Press; 1969:21-132.

53. Saitou N, Nei M: **The neighbor-joining method: a new method for reconstructing phylogenetic trees.** *Mol Biol Evol.* 1986, **4:** 406-425.

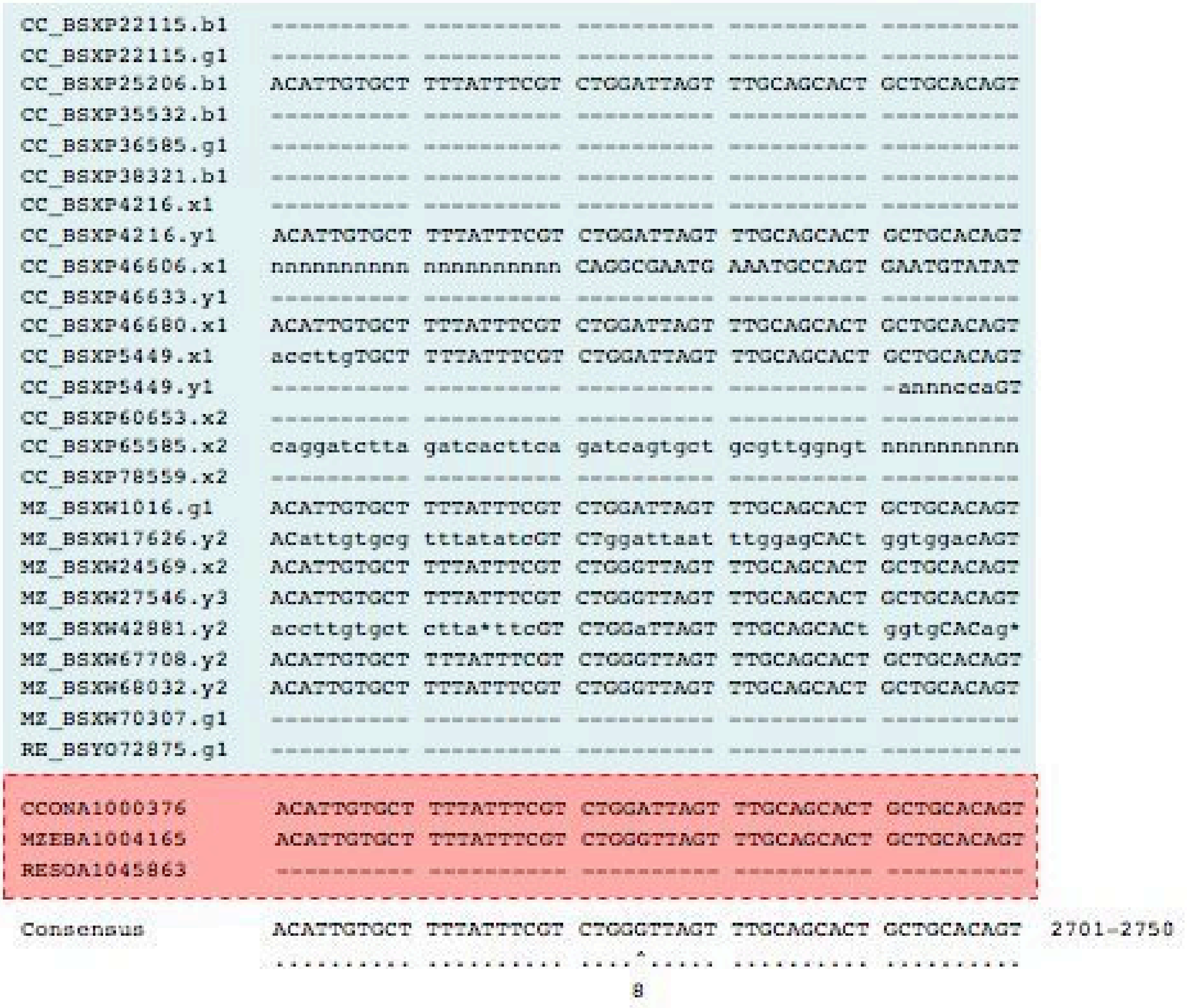
54. Felsenstein J: **PHYLIP (Phylogeny Inference Package), version 3.67.** Department of Genetics, University of Washington, Seattle; 2007.

Figure 1.

A.



B.



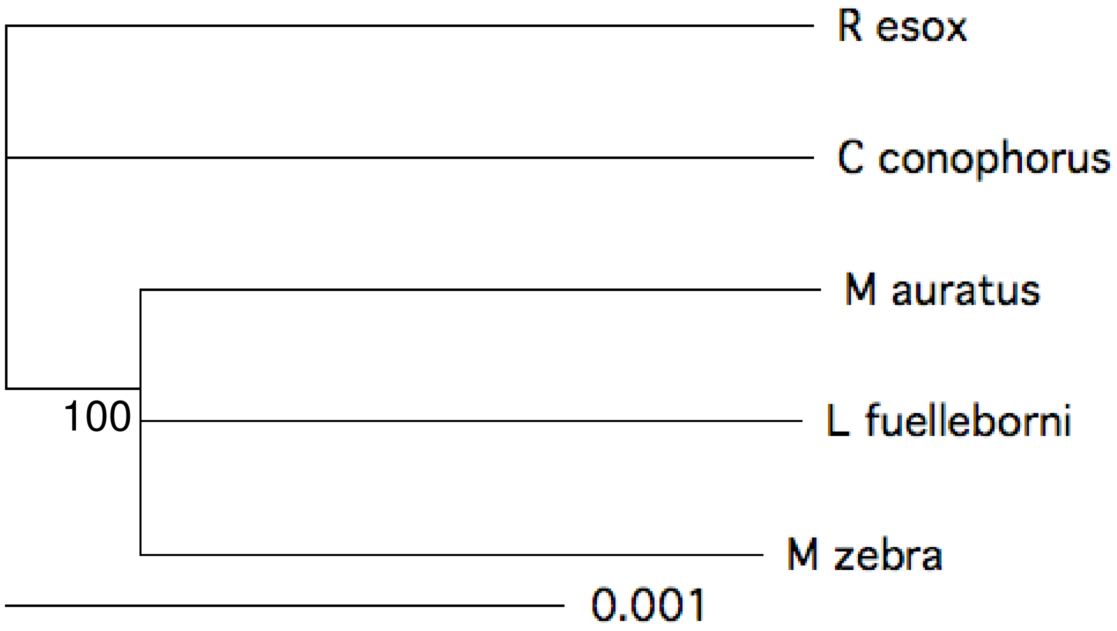
```
CC_BSXP22115.b1      ---------- ---------- ---------- ---------- ----------
CC_BSXP22115.g1      ---------- ---------- ---------- ---------- ----------
CC_BSXP25206.b1      ACATTGTGCT TTTATTTCGT CTGGATTAGT TTGCAGCACT GCTGCACAGT
CC_BSXP35532.b1      ---------- ---------- ---------- ---------- ----------
CC_BSXP36585.g1      ---------- ---------- ---------- ---------- ----------
CC_BSXP38321.b1      ---------- ---------- ---------- ---------- ----------
CC_BSXP4216.x1       ---------- ---------- ---------- ---------- ----------
CC_BSXP4216.y1       ACATTGTGCT TTTATTTCGT CTGGATTAGT TTGCAGCACT GCTGCACAGT
CC_BSXP46606.x1      nnnnnnnnnn nnnnnnnnnn CAGGCGAATG AAATGCCAGT GAATGTATAT
CC_BSXP46633.y1      ---------- ---------- ---------- ---------- ----------
CC_BSXP46680.x1      ACATTGTGCT TTTATTTCGT CTGGATTAGT TTGCAGCACT GCTGCACAGT
CC_BSXP5449.x1       accttgTGCT TTTATTTCGT CTGGATTAGT TTGCAGCACT GCTGCACAGT
CC_BSXP5449.y1       ---------- ---------- ---------- ---------- -annnccaGT
CC_BSXP60653.x2      ---------- ---------- ---------- ---------- ----------
CC_BSXP65585.x2      caggatctta gatcacttca gatcagtgct gcgttggngt nnnnnnnnnn
CC_BSXP78559.x2      ---------- ---------- ---------- ---------- ----------
MZ_BSXW1016.g1       ACATTGTGCT TTTATTTCGT CTGGATTAGT TTGCAGCACT GCTGCACAGT
MZ_BSXW17626.y2      ACattgtgcg tttatatcGT CTggattaat ttggagCACt ggtggacAGT
MZ_BSXW24569.x2      ACATTGTGCT TTTATTTCGT CTGGGTTAGT TTGCAGCACT GCTGCACAGT
MZ_BSXW27546.y3      ACATTGTGCT TTTATTTCGT CTGGGTTAGT TTGCAGCACT GCTGCACAGT
MZ_BSXW42881.y2      accttgtgct ctta*ttcGT CTGGaTTAGT TTGCAGCACt ggtgCACag*
MZ_BSXW67708.y2      ACATTGTGCT TTTATTTCGT CTGGGTTAGT TTGCAGCACT GCTGCACAGT
MZ_BSXW68032.y2      ACATTGTGCT TTTATTTCGT CTGGGTTAGT TTGCAGCACT GCTGCACAGT
MZ_BSXW70307.g1      ---------- ---------- ---------- ---------- ----------
RE_BSY072875.g1      ---------- ---------- ---------- ---------- ----------

CCONA1000376         ACATTGTGCT TTTATTTCGT CTGGATTAGT TTGCAGCACT GCTGCACAGT
MZEBA1004165         ACATTGTGCT TTTATTTCGT CTGGGTTAGT TTGCAGCACT GCTGCACAGT
RESOA1045863         ---------- ---------- ---------- ---------- ----------

Consensus            ACATTGTGCT TTTATTTCGT CTGGGTTAGT TTGCAGCACT GCTGCACAGT  2701-2750
                                               ^
                     .......... .......... .......... .......... ..........
                                              8
```

Figure 2.

	RE	MZ	MA	LF
CC	0.00287	0.00281	0.00291	0.00284
LF	0.00288	0.00229	0.00241	
MA	0.00286	0.00232		
MZ	0.00280			



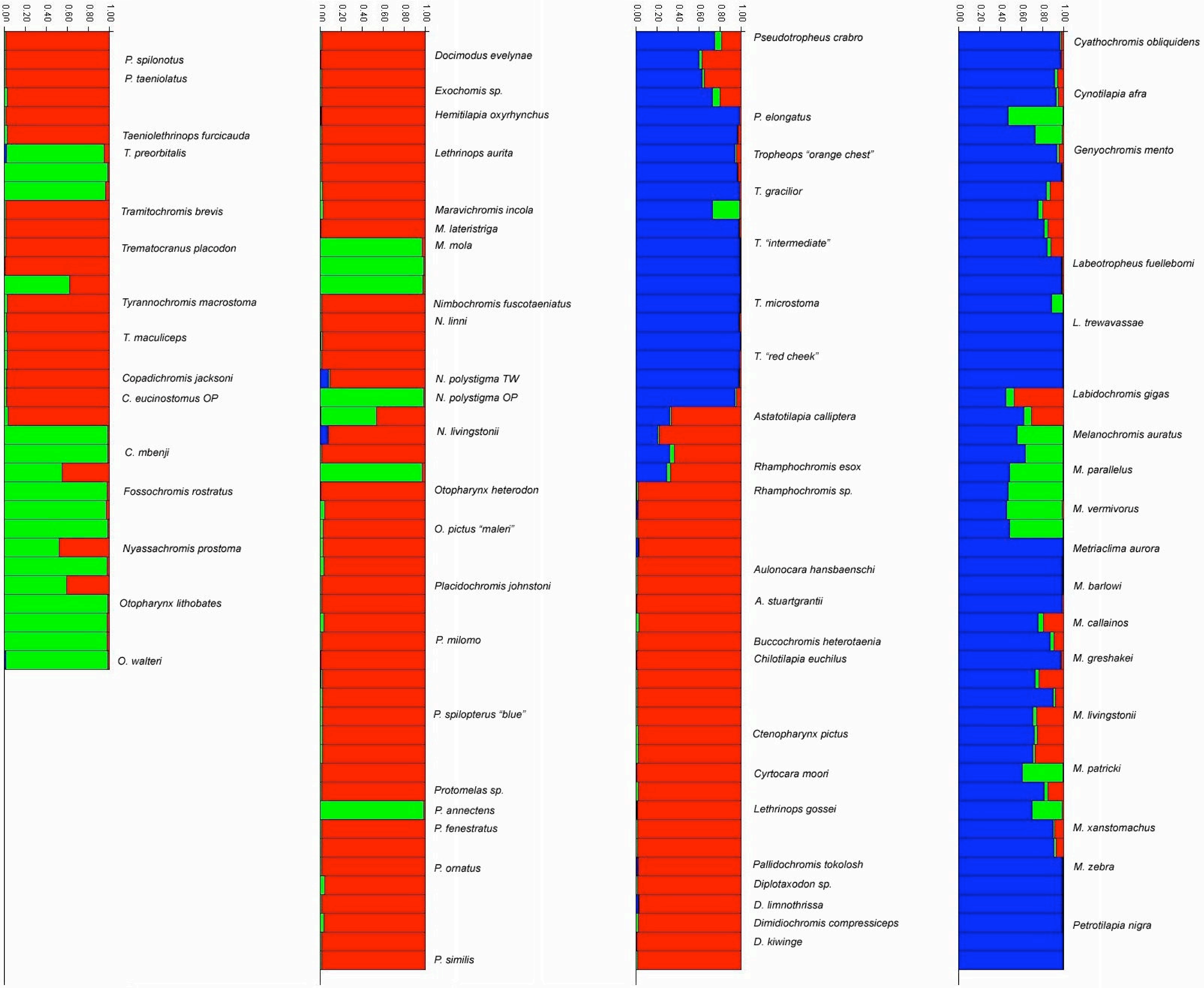

0.00
0.20
0.40
0.60
0.80
1.00
P. spilonotus
P. taeniolatus
Taeniolethrinops furcicauda
T. preorbitalis
Tramitochromis brevis
Trematocranus placodon
Tyrannochromis macrostoma
T. maculiceps
Copadichromis jacksoni
C. eucinostomus OP
C. mbenji
Fossochromis rostratus
Nyassachromis prostoma
Otopharynx lithobates
O. walteri
0.00
0.20
0.40
0.60
0.80
1.00
Docimodus evelynae
Exochomis sp.
Hemitilapia oxyrhynchus
Lethrinops aurita
Maravichromis incola
M. lateristriga
M. mola
Nimbochromis fuscotaeniatus
N. linni
N. polystigma TW
N. polystigma OP
N. livingstonii
Otopharynx heterodon
O. pictus "maleri"
Placidochromis johnstoni
P. milomo
P. spilopterus "blue"
Protomelas sp.
P. annectens
P. fenestratus
P. ornatus
P. similis
0.00
0.20
0.40
0.60
0.80
1.00
Pseudotropheus crabro
P. elongatus
Tropheops "orange chest"
T. gracilior
T. "intermediate"
T. microstoma
T. "red cheek"
Astatotilapia calliptera
Rhamphochromis esox
Rhamphochromis sp.
Aulonocara hansbaenschi
A. stuartgrantii
Buccochromis heterotaenia
Chilotilapia euchilus
Ctenopharynx pictus
Cyrtocara moori
Lethrinops gossei
Pallidochromis tokolosh
Diplotaxodon sp.
D. limnothrissa
Dimidiochromis compressiceps
D. kiwinge
0.00
0.20
0.40
0.60
0.80
1.00
Cyathochromis obliquidens
Cynotilapia afra
Genyochromis mento
Labeotropheus fuelleborni
L. trewavassae
Labidochromis gigas
Melanochromis auratus
M. parallelus
M. vermivorus
Metriaclima aurora
M. barlowi
M. callainos
M. greshakei
M. livingstonii
M. patricki
M. xanstomachus
M. zebra
Petrotilapia nigra

Figure 3

Additional files provided with this submission:

Additional file 1: cichlidsnps_suppltable1_111607.doc, 38K
http://genomebiology.com/imedia/1261566269172259/supp1.doc
Additional file 2: cichlidsnps_suppltable2_111607.xls, 1775K
http://genomebiology.com/imedia/1243272635172259/supp2.xls
Additional file 3: cichlidsnps_suppltable3_111607.xls, 2908K
http://genomebiology.com/imedia/6562457741722592/supp3.xls
Additional file 4: cichlidsnps_suppltable4_111607.xls, 1640K
http://genomebiology.com/imedia/2648011871722592/supp4.xls