MOLECULAR ECOLOGY

Molecular Ecology (2010), **19** (Suppl. 1), 197–211 doi: 10.1111/j.1365-294X.2009.04488.x

Rapid evolution and selection inferred from the transcriptomes of sympatric crater lake cichlid fishes

K. R. ELMER, S. FAN, H. M. GUNTER, J. C. JONES, S. BOEKHOFF, S. KURAKU and A. MEYER

Lehrstuhl für Zoologie und Evolutionsbiologie, Department of Biology, University of Konstanz, Universitätstrasse 10, 78457 Konstanz, Germany

Abstract

Crater lakes provide a natural laboratory to study speciation of cichlid fishes by ecological divergence. Up to now, there has been a dearth of transcriptomic and genomic information that would aid in understanding the molecular basis of the phenotypic differentiation between young species. We used next-generation sequencing (Roche 454 massively parallel pyrosequencing) to characterize the diversity of expressed sequence tags between ecologically divergent, endemic and sympatric species of cichlid fishes from crater lake Apoyo, Nicaragua: benthic *Amphilophus astorquii* and limnetic *Amphilophus zaliosus*. We obtained 24 174 *A. astorquii* and 21 382 *A. zaliosus* high-quality expressed sequence tag contigs, of which 13 106 pairs are orthologous between species. Based on the ratio of nonsynonymous to synonymous substitutions, we identified six sequences exhibiting signals of strong diversifying selection ($K_a/K_s > 1$). These included genes involved in biosynthesis, metabolic processes and development. This transcriptome sequence variation may be reflective of natural selection acting on the genomes of these young, sympatric sister species. Based on *K*s ratios and p-distances between 3′-untranslated regions (UTRs) calibrated to previously published species divergence times, we estimated a neutral transcriptome-wide substitutional mutation rate of $\sim 1.25 \times 10^{-6}$ per site per year. We conclude that next-generation sequencing technologies allow us to infer natural selection acting to diversify the genomes of young species, such as crater lake cichlids, with much greater scope than previously possible.

Keywords: adaptive radiation, comparative genomics, expressed sequence tags, natural selection, pyrosequencing, Roche 454 GS FLX, substitutional mutation rate, Nicaragua

Received 15 July 2009; revision received 6 October 2009; accepted 9 October 2009

Introduction

Cichlid fishes are one of the most species-rich vertebrate families and much of this richness is reflected in morphological adaptations related to trophic niche, such as body shape, mouth and jaw form (Rüber *et al.* 1999; Clabaut *et al.* 2007; Salzburger 2009). The Midas cichlids in Nicaragua have evolved rapidly into different trophic niches in young crater lakes (Barlow & Munsey 1976; Wilson *et al.* 2000; Vivas & McKaye 2001; Stauffer & McKaye 2002; Barluenga *et al.* 2006; Stauffer *et al.* 2008). In crater lake Apoyo, such eco-morphological speciation has occurred in sympatry: a trophically and genetically distinct endemic limnetic species, *Amphilophus zaliosus*, evolved from a benthic or generalist ancestor in the past 10 000 years (Barlow & Munsey 1976; Barluenga *et al.* 2006) (Fig. 1). *Amphilophus astorquii*, a recently described benthic species, is also strictly endemic to crater lake Apoyo (Stauffer *et al.* 2008), where it is the most abundant species in the lake (McCrary & López 2008). Disruptive natural selection within the crater lake environment is likely to have driven species' divergent ecologies and morphologies (Barlow & Munsey 1976; Parsons *et al.* 2003; Barluenga *et al.* 2006; Elmer *et al.* in press), resulting in reproductive isolation (Baylis 1976) and population differentiation at putatively neutral genetic regions

Correspondence: Axel Meyer, Fax: +49 7531 88 3018;
E-mail: axel.meyer@uni-konstanz.de

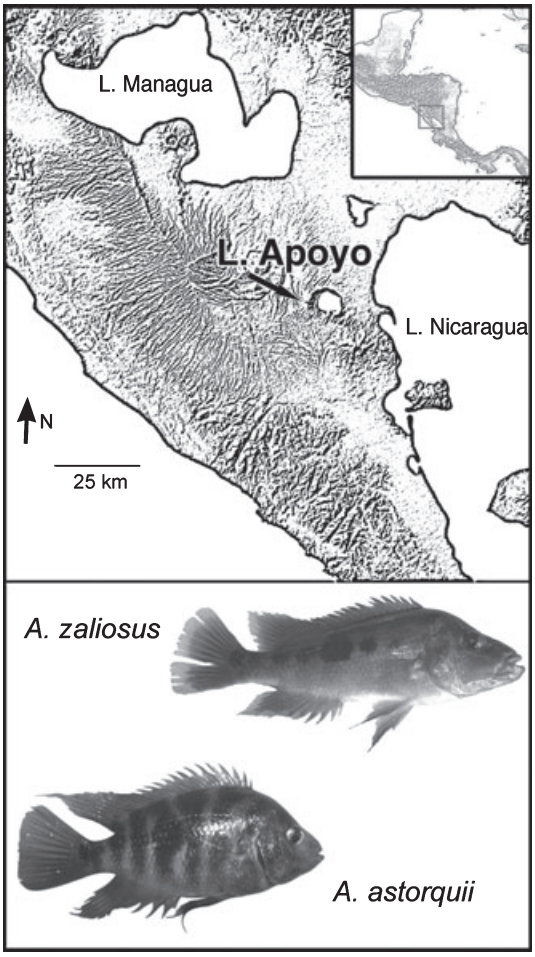

Fig. 1 Lake Apoyo is a crater lake in western Nicaragua and has no water connection with neighbouring crater or great lakes. *Amphilophus zaliosus* is a limnetic endemic species that evolved in sympatry in Lake Apoyo (lower above). *Amphilophus astorquii* is an endemic benthic species (lower below).

(microsatellites, amplified fragment length polymorphisms, mitochondrial DNA) (Barluenga *et al.* 2006). Given the Lake Apoyo Midas cichlid species flock's genetic monophyly (Wilson *et al.* 2000; Barluenga & Meyer 2004; Barluenga *et al.* 2006; Bunje *et al.* 2007), the species' strict endemism, and trophic differentiation (Stauffer *et al.* 2008), it is highly probable that not only *A. zaliosus* (Barlow & Munsey 1976; Barluenga *et al.* 2006) but also *A. astorquii* arose in lake Apoyo by sympatric speciation (Stauffer *et al.* 2008). Despite the established ecological and neutral genetic differences, we lack estimates of gene sequence diversity between the benthic and limnetic species. Further, we have no information about how divergent natural selection may have affected these species' genomes within the very short evolutionary time span since their divergence from a common ancestor.

Emerging techniques based on 'next generation' sequencing or massively parallel sequencing, such as low-pass shotgun genome sequencing and expressed sequence tag (EST) analyses, are proving to be valuable additions to evolutionary and ecological research (Ellegren 2008; Rokas & Abbot 2009). Research areas such as population genetics (Lynch *et al.* 2008), experimental evolution (Shendure *et al.* 2005) and phylogenetics (Moore *et al.* 2006, 2007) have successfully synthesized genome data to address biologically meaningful questions. The spearhead for research on the genomic bases of species differences is work on human genome comparisons, which seek to identify the features that distinguish us from our primate relatives and to quantify the variability between human populations. For example, a comparison between the human and chimp genomes indicated that they are 98.4% similar (Chen & Li 2001), and that genes related to immune functions, spermatogenesis, olfaction and sensory perception are probably under positive selection in the human lineage (Bustamante *et al.* 2005; Nielsen *et al.* 2007). This human research has driven the development of cost-effective techniques, which have opened the door for genome research on nonmodel organisms: the 'new frontiers' of genomics research (Collins *et al.* 2003). There are specific challenges that are encountered when using massively parallel sequencing technologies on nonmodel organisms because a reference genome is rarely available. In some cases, a close relative provides a scaffold for read assembly: for example, having a sequenced honeybee genome facilitated wasp transcriptome research (Toth *et al.* 2007), the silkworm genome aided analysis of Glanville fritillary butterfly ESTs (Vera *et al.* 2008) and chicken and zebra finch genomes inform transcriptome research on non-model bird species (Kunstner *et al.* 2010; Wolf *et al.* 2010). Sufficiently long sequence reads from massively parallelized sequencing can compensate for the lack of a reference genome so 454 pyrosequencing, with currently the longest available read lengths, is the platform favoured for such scenarios (Rokas & Abbot 2009). In the context of our research, a completely sequenced cichlid genome is underway (International Cichlid Genome Consortium 2006) but is not yet complete.

Expressed sequence tags represent a sample of the spatiotemporally expressed genome: the transcriptome. EST studies can be used as an entry into gene expression and comparative genome-level questions in nonmodel organisms when other genomic resources, such as a sequenced genome, are not yet developed (Bouck & Vision 2007; Hudson 2007). EST studies – one of the most cost-effective methods for gene discovery (Bouck & Vision 2007) – are made even more robust and efficient using massively parallelized sequencing. This has

eliminated the need for cloning ESTs, which introduces bias, and has greatly increased the quantity of data that can be generated in a short time at a reduced cost compared with traditional Sanger sequencing of cDNA libraries (Weber *et al.* 2007; Wheat, in press). Parallelized sequencing of transcriptomes allows us to identify candidate genes without imposing strong *a priori* expectations or biases (although most studies restrict their starting material to pertinent tissues and/or developmental stages when resources are limited). Implicit in this approach to identify transcriptome differences between species is the expectation that gene (or transcript) sequence differences may be relevant to some interspecific phenotypic variation.

In this study, we use massively parallelized pyrosequencing (454 GS FLX) to characterize the transcriptome sequence diversity between two young, endemic and ecologically divergent sympatric species of Midas cichlid fish from the neotropical crater lake Apoyo, Nicaragua. *A. zaliosus* is an open water, elongate limnetic species. *A. astorquii* is a high-bodied, short benthic species. Both species breed in the littoral zone during the same breeding season, with *A. zaliosus* breeding in solitary pairs and *A. astorquii* in colonies of pairs. We aim to identify interspecific EST sequence variation and infer orthologous genes that may be showing signs of diversifying natural selection (a non-neutral rate of synonymous and nonsynonymous substitutions between sequence pairs). We identify transcriptome sequence variation that reflects the impact of natural selection on the genome. Using the neutral substitutional mutation rate inferred from K_s and the similarities in 3′-untranslated regions (UTR), we estimate a transcriptome-wide substitutional mutation rate for these neotropical cichlids. By using pyrosequencing technology, we infer mutation rates and the effects of natural selection across the genome and transcriptome with greater speed and scope than previously possible.

Materials and methods

Generating samples

Intraspecific crosses were established between pairs of *Amphilophus astorquii* and *Amphilophus zaliosus*. These adults have been laboratory-reared under common conditions since they were collected in Lake Apoyo as fry 2–4 years before. When a clutch was successfully produced, three time points were sampled from each species: day of hatching (1 dph), 1 week post-hatching (1 wph) and 1 month post-hatching (1 mph). Samples were collected in a standardized manner and time of day. Fry and juvenile fish were killed following approved protocols and placed in RNALater (Qiagen). These were held at 4 °C for less than 1 month before being stored at −20 °C. A portion of each sample not stored in RNALater was fixed in 4% paraformaldehyde overnight, stepped into methanol, and then stored at −20 °C to provide an archive of the stage-specific morphologies.

RNA extraction

RNA extractions were performed simultaneously. For each species, sample sizes were: 6 individuals of 1 dph, 10 individuals of 1 wph, and 2 individuals of 1 mph. Fewer individuals were included in 1 dph than 1 wph because the high quantity of yolk in the 1 dph samples was found to cause poor quality and quantity of total RNA. Only the head (cut immediately behind the gill cover) was used for 1 mph samples to avoid contaminating the sample with gut flora and fauna. After removing RNALater, samples were homogenized in 1 mL of Trizol (Invitrogen) in an MP Biosciences homogenizer at intensity 5.0 for 20 s. RNA extraction was performed using the manufacturer's protocol and re-precipitated for 3 h with one volume of 4M LiCl. Pellets were recovered by centrifugation and dissolved in 20 μL pH 8.0 diethylpyrocarbonate water (DEPC). The quantity and quality of total RNA was assessed by spectrophotometry and gel electrophoresis. Between 1 and 2 μg of each sample was requested for commercial normalized library construction, and equal quantities of RNA from each stage were pooled per species.

Normalized library development

We commissioned a 3′-fragment normalized cDNA library to be constructed by a third-party service provider (Eurofins MWG GmbH, Ebersberg, Germany). Briefly, from total RNA, first-strand cDNA was synthesized using reverse transcriptase and an oligo(dT)-adapter primer. Second-strand synthesis was performed with a N6 random adapter primer. cDNAs were then amplified with 17 (*A. zaliosus*) or 18 (*A. astorquii*) cycles of long and accurate polymerase chain reaction (PCR) (Barnes 1994). Libraries were normalized by hydroxylapatite chromatography and the ss-cDNA was then amplified by PCR. cDNA was size selected for 450–550 bp including the 5′- and 3′-454 and cDNA adapters.

454 Sequencing and assembly quality control

The normalized cDNA library was sequenced in one GS FLX (Roche 454) Standard Chemistry run (half a plate per species at equimolar concentrations) by Eurofins MWG. Reads were assembled using the Eurofins MWG in-house bioinformatics pipeline. Adapter sequences

were clipped and contigs were assembled using MIRA 2.9.15 (Chevreux *et al.* 2004) based on 40 bp overlap, 90% homology and a minimum of five reads deep on average. Singleton raw reads were excluded.

We subjected all contigs to an extensive quality control procedure. First, sequences were screened for contamination by BLASTn searches (E-value E-20) against *Escherichia coli* genome and human and mouse EST databases (downloaded December 2008). Two *E. coli* contamination sequences were identified in the *A. astorquii* pool and excluded. Second, low-quality bases were masked using an in-house Perl script (S. Fan) given a quality score threshold of $Q > 20$. Third, contigs with interspersed repeats and low-complexity DNA sequences were excluded using an in-house script (S. Fan) parsing the results of RepeatMasker (version 3.2.6 with repeat library 20090120) (Smit *et al.* 1996), since such sequences would impede orthologous EST identification. Contigs <200 bp long were excluded from further analyses.

Transcriptome functional annotation

Functional annotation was performed online using Blast2GO (Version 2.3.4) (Conesa *et al.* 2005; Götz *et al.* 2008), which performs a BLASTX search against the nonredundant database on NCBI (default parameters were used). Annotated accession numbers and Gene Ontology (GO) (The Gene Ontology Consortium 2000) numbers were derived from NCBI QBLAST (Altschul *et al.* 1997) based on an E-value ≤1E-5 and a high-scoring segment pair cut-off greater than 33. The annotation procedure was conducted using the following parameters: a pre-E-value-Hit-Filter of 10^{-6}, a pro-Similarity-Hit-Filter of 15, an annotation cut-off of 55, and a GO weight of 5.

Identifying orthologous ESTs

We used the bidirectional best hit method in BLAST with a bit score threshold of >300 to identify ESTs that are putatively orthologous between the two species. Bidirectional best hit has been found to out-perform more complex orthology identification algorithms (Altenhoff & Dessimoz 2009). Our bidirectional best hit threshold ensures that the alignment of two ESTs is longer than 150 bp.

Predicting the open reading frame and the untranslated region

Open reading frames (ORF) for the putatively orthologous ESTs were determined by BLASTX (NCBI blast version, 2.2.19) (Altschul *et al.* 1997) against all known vertebrate proteins from the Universal Protein Resource (The UniProt Consortium 2008) and protein data sets for five teleost fishes (fugu, medaka, green spotted pufferfish, stickleback and zebrafish) in the Ensembl database (Hubbard *et al.* 2005) (Ensembl 52) using a threshold of <1E-5. If both orthologous ESTs could be annotated, the coding regions were extracted according to the BLASTX results. The coding sequences were aligned by ClustalW version 2.0 (Larkin *et al.* 2007).

The 3′-untranslated region (UTR) of each contig was identified based on the results of the ORF prediction. We searched downstream of the coding region to identify the stop codon (TAG, TAA or TGA). If the number of base pairs between the stop codon and end of the coding region were divisible by three (i.e. matched a reading frame by being the length of an amino acid), then downstream of the stop codon was considered a 'true 3′-UTR'. If the number of base pairs between the coding region and the stop codon was not divisible by three, then downstream of the ORF was considered a 'pseudo-UTR' and excluded from further analyses.

Estimating substitution rates

We estimated the rate of nonsynonymous substitutions per nonsynonymous site (K_a) to the number of synonymous substitutions per synonymous site (K_s) between putatively orthologous coding regions using a maximum-likelihood method (Yang & Nielsen 2000) implemented by yn00 in the PAML toolkit (vers. 4.0) (Yang 2007). Orthologous ESTs with a K_s rate >0.1 were excluded from further analyses to avoid analysing paralogous genes (Bustamante *et al.* 2005).

Estimating the overall substitutional mutation rate

We estimated an overall substitution rate for the cichlid genome based on divergence between orthologous EST pairs (entire EST, including coding region and UTRs > 50 bp long) and synonymous mutations calibrated with a maximum age of crater Lake Apoyo (Kutterolf *et al.* 2007). Only UTRs contiguous with orthologous coding regions were used in distance calculations to avoid including artefacts of assembly. The rate (r) (in substitutions/site/year) is calculated from the mean genetic distance between sequences (d) divided by the divergence time between two species ($2t$). d for coding regions is based on the K_s rate, since under the neutral theory of evolution K_s should be proportional to the neutral mutation rate (Hurst 2002). d for UTRs was estimated by a Jukes-Cantor (Jukes & Cantor 1969) corrected pairwise distance.

Results

Sequencing and assembly quality control

We received a total of 114 Mb from one run, approximately evenly represented in both species (Table 1). Sequences are deposited in the NCBI Short Read Archive (Accession no. SRA009759.2). The average read length was just over 200 bp (Table 1). This is shorter than the 220–270 average read length expected from the GS FLX technology and may be because of the 3′-library construction method. Our total number of base pairs meets Roche 454 expectations of 100 Mb per run for GS FLX standard chemistry. Raw reads were assembled into 57 566 contigs. Fewer than 100 hits per pool could be attributed to the mitochondrial genome (from neotropical cichlids: NCBI Accession nos NC_009058, NC_011168). After quality control (see Methods and materials), our sample consisted of 24 174 *Amphilophus astorquii* and 21 382 *Amphilophus zaliosus* 'high-quality ESTs' ranging in length from 200 to 1277 bp. These sequences were used for further analysis.

A total of 2289 *A. astorquii* and 2119 *A. zaliosus* ESTs showed homology (i.e. significant e-values) with known proteins from the vertebrate protein database (Universal Protein Resource) and five fish protein databases (Ensembl). This represents about 10% of the ESTs being successfully annotated, a proportion that is less than the 20% to 40% of ESTs often annotated from a traditional Sanger sequenced EST library (e.g. Cerda *et al.* 2008; Salzburger *et al.* 2008) but similar in absolute numbers. By proportion our annotation success is lower because 454 reads tend to be biased towards the 3′-transcript end (Shin *et al.* 2008) and are shorter than traditional Sanger sequences. Also, we used a 3′-fragment extension library protocol that should maximize depth but at the cost of 5′ ends, which makes annotation more difficult (G. Gradl, personal communication).

Functional annotation

Of our data, 3152 (13%) of the *A. astorquii* and 2673 (12%) of the *A. zaliosus* contigs were annotated with an inferred biological function based on currently known proteins in the NCBI nonredundant protein database (BLASTX). Approximately equal numbers of the EST sequences for *A. astorquii* and *A. zaliosus* had GO resource assignments relating to three major divisions. The first, 'biological process', refers to the 'biological objective to which the gene or gene product contributes' (The Gene Ontology Consortium 2000). Within the function of 'biological process', 15 categories were identified and these were perfectly paired between species. The two most abundant categories were: (i) 'cellular and metabolic processes', to which 46% of both species' ESTs were dedicated (2810 *A. astorquii* and 2705 *A. zaliosus* ESTs); and (ii) 'development process', to which 10% of ESTs were dedicated (627 *A. astorquii* and 593 *A. zaliosus* ESTs) (Fig. 2). The second major division is 'molecular function', which refers to some biochemical activity that is performed by the gene, without a temporal or spatial context (The Gene Ontology Consortium 2000). EST coverage of this division was similar

Table 1 Sequencing coverage was approximately equal in both species, suggesting that there was no bias towards any particular pool of RNA

	Amphilophus astorquii	*Amphilophus zaliosus*	Total
Total number of reads	300 610	262 494	563 104
Total number of bases	60 732 547	54 132 058	114 864 605
Average read length	202	206	
Assembly results			
Number assembled	231 293	199 628	
Number too short and excluded	13 749	11 567	
Number of all contigs	30 569	26 997	57 566
Total number of bases	8 143 429	7 167 699	15 311 128
'High quality' ESTs			
Total number of contigs	24 174	21 382	45 556
Average length	300	299	
±SD	79	76	
Median length	275	277	
Maximum length	1 277	1 271	

'High quality' ESTs are those used for subsequent analyses. Total represents the two species combined. Numbers of bases and read length are trimmed of tags and low quality bases. EST, expressed sequence tags.

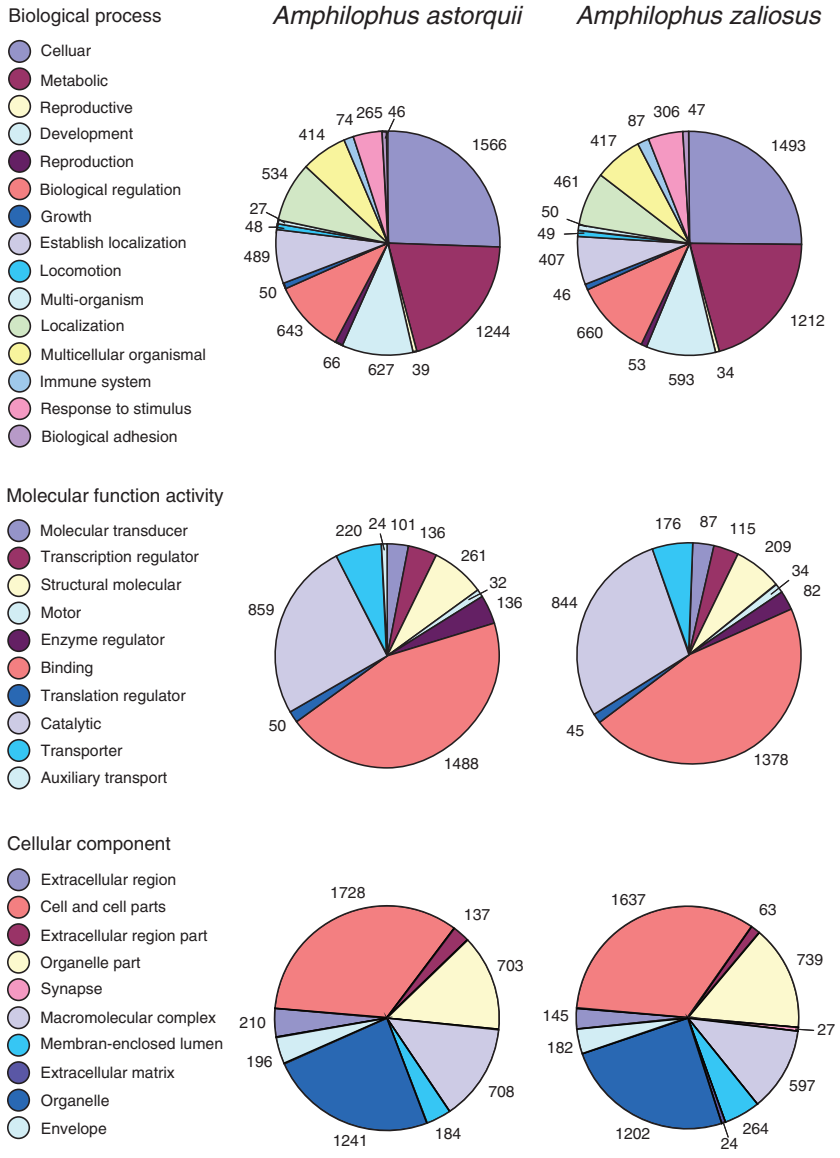

Fig. 2 Blast2GO assignment for 3152 *Amphilophus astorquii* and 2673 *Amphilophus zaliosus* ESTs. The proportion of ESTs assigned to different categories is approximately equal.

between species: 10 categories of 'molecular function' were found in *A. astorquii* and nine categories in *A. zaliosus* (genes corresponding to auxiliary transport protein activity function could only be found in the *A. astorquii* transcriptome) (Fig. 2). Of these ESTs ascribed to 'molecular function', most *A. astorquii* (71%, 2347 sequences) and *A. zaliosus* (75%, 2222 sequences) ESTs were dedicated to binding functions and catalytic activity. The third division is 'cellular component', which describes the sub-cellular location where a gene product is active (The Gene Ontology Consortium 2000). Again, coverage is similar between species: nine categories were found in the *A. astorquii* transcriptome and 11 categories were found in the *A. zaliosus* transcriptome. Gene products were mainly expressed intracellularly (*A. astorquii*: 3456 sequences or 51%; *A. zaliosus*: 3274 sequences or 50%) or in the organelle (*A. astorquii* 1241 sequences or 18%; *A. zaliosus*: 1202 sequences or 18%). The matched proportion of GO categories between *A. astorquii* and *A. zaliosus* suggests that our library and 454 sequencing covered both species' transcriptomes equally.

Orthologous EST identification

We identified 13 106 pairs of ESTs that are putatively orthologous between the two species (hereafter referred to as 'orthologous ESTs'). The median length of sequence shared by the orthologous ESTs (i.e. alignment length) is 264 bp, ranging from 153 to 767 bp. A total of 1721 pairs of orthologous ESTs matched to the ORFs of known or unknown proteins.

Untranslated region identification

The untranslated region of each orthologous EST was identified based on the predicted coding region. Thirty-three pairs (median length 78 bp, ranging from 51 to 153 bp) of UTRs that are informative for divergence were found in the orthologous ESTs.

Estimated Ka/Ks

Based on a data set of 1721 pairs of ESTs that were orthologous and had an ORF that could be predicted (criterion 1E-5), divergence was sufficiently high for 44 ESTs (3%) of which both a K_a and a K_s rate could be calculated. Of these, six orthologous ESTs have a $K_a/K_s > 1$ and eight orthologous ESTs have a K_a/K_s between 0.5 and 1 (Fig. 3). For the remainder of the orthologous ESTs, we could calculate either only K_a (175 orthologous ESTs, 10%), only K_s (103 orthologous ESTs, 6%), or the orthologous ESTs were identical (1399 or 81%), making a ratio incalculable. $K_a/K_s > 1$ suggests that strong positive selection has acted to change the protein DNA sequence (Yang & Bielawski 2000) while K_a/K_s above 0.5 is a less conservative cut-off that has also proven useful for identifying genes under positive selection (Swanson *et al.* 2004). EST pairs with $K_a/K_s > 1$ function in biosynthetic and metabolic processes, brain development and cognition, response to hormone stimuli, and nervous system development. Those ESTs with K_a/K_s between 0.5 and 1 related to metabolic processes, tissue, blood and hormone regulation and regeneration (Table 2).

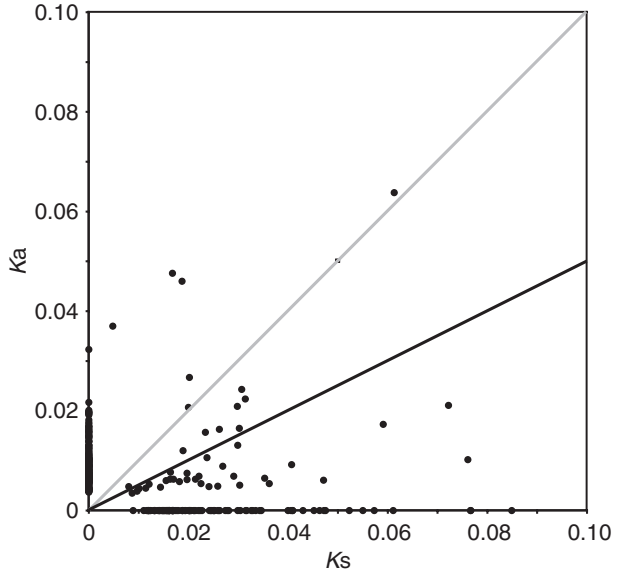

Fig. 3 Distribution of K_a and K_s. A nonzero K_a and a K_s ratio could be calculated for 44 ESTs. ESTs with $Ka/Ks > 1$ fall above the grey line while ESTs with $K_a/K_s = 0.5–1$ fall between the black and grey lines.

Substitution rate estimation

The average K_s rate for 147 orthologous ESTs is 0.0250 ± 0.015 (mean ± SD). The average divergence between orthologous and informative UTRs is 0.0252 ± 0.020 (Fig. 4). Substitutions in the synonymous sites and 3′-UTR are putatively neutral, especially among very closely related taxa (Hurst 2002) (but see Hellmann *et al.* 2003, who find 5′-UTR in humans may be under positive selection). The similarity of the two means strongly suggests that our coding region and UTR analyses are valid and represent equal and approximately neutral evolutionary change.

Based on a species divergence time of 10 000 years ago estimated from mitochondrial DNA (Barluenga *et al.* 2006), we used the substitution rate estimated above to calculate a transcriptome-wide substitution rate of 1.25×10^{-6} per site per year. The geological age of Lake Apoyo is 23 890 ± 120 years (Kutterolf *et al.* 2007). Using that as a maximum species divergence time, we infer a minimum transcriptome-wide substitution rate of 6.3×10^{-7} mutations per site per year (calibrated to 20 000 years).

Discussion

Transcriptome variation in the Midas cichlid species complex

We have identified genes under balancing and positive selection in the extremely young crater lake cichlid species *Amphilophus astorquii* and *Amphilophus zaliosus*. These species are model systems for understanding the ecology and evolution of adaptive radiations and sympatric speciation (Barluenga *et al.* 2006; Elmer *et al.* in press). *A. astorquii* is endemic to crater lake Apoyo (Stauffer *et al.* 2008), which houses a monophyletic Midas cichlid species flock (Bunje *et al.* 2007). Thus, although the ecology and evolutionary history of the recently described *A. astorquii* have not yet been extensively studied (but see Elmer *et al.* in press; Stauffer *et al.* 2008; Oldfield 2009) it is very likely that, like *A. zaliosus* (Barluenga *et al.* 2006), *A. astorquii* arose by sympatric speciation (Stauffer *et al.* 2008). Because of the young ages of these species, we anticipated that genetic differences between species would be very small. This expectation was borne out and in this initial screen we find only 14 candidate genes that show signs of positive selection while most of the transcriptome is either identical between species or was only sampled in one species. This is in agreement with previous research

Table 2 Ratio of nonsynonymous (K_a) to synonymous (K_s) substitutions, number of polymorphisms (SNPs), bit scores, E-values, Gene Ontology (F, molecular function, P, biological process, C, cellular components), hit sequence name, the species of origin for this hit and the accession identification numbers for homologues in other species of the 14 candidate genes (ESTs) showing a signal of evolution by diversifying natural selection

K_a/K_s	Number of SNPs	*Amphilophus astorquii* Bit score	*Amphilophus zaliosus* Bit score	*Amphilophus astorquii* E-value	*Amphilophus zaliosus* E-value	Gene ontology	Hit sequence name	Species of origin	Accession IDs
7.682	4	88.2	148	3.00E-17	2.00E-35	F: isomerase activity; P: biosynthetic process	Phenazine biosynthesis-like domain-containing protein 2	*Salmo salar*	tr|B5X8N7|B5X8N7_SALSA
2.836	5	82.8	127	1.00E-15	5.00E-29	P: amino acid metabolic process; P: ATP metabolic process; C: lysosomal membrane; C: integral to membrane; F: L-cystine transmembrane transporter activity; P: brain development; P: cognition; P: L-cystine transport; P: glutathione metabolic process; C: late endosome; C: early endosome	Cystinosin	*Oryzias latipes*	ENSORLP00000003583
2.453	7	70.1	120	7.00E-12	6.00E-27	Not available	Novel protein coding	*Oryzias latipes*	ENSORLP00000003819
1.320	4	143	125	7.00E-34	1.00E-28	Not available	Conserved domain	*Gasterosteus aculeatus*	ENSGACP00000016806
1.040	8	87	131	6.00E-17	3.00E-30	C: cytoplasm; P: response to hormone stimulus; P: epithelial cell differentiation; C: polysome; P: nervous system development; F: poly(U) binding; C: nucleus; F: nucleotide binding	RNA-binding protein Musashi homologue 1	*Oryzias latipes*	ENSORLP00000024802
1.035	4	92	58.2	2.00E-18	3.00E-08	P: cell redox homeostasis; P: transport	Thioredoxin domain	*Bombina orientalis*	tr|B6VFL5|B6VFL5_BOMOR

Table 2 *Continued*

K_a/K_s	Number of SNPs	*Amphilophus astorquii* Bit score	*Amphilophus zaliosus* Bit score	*Amphilophus astorquii* E-value	*Amphilophus zaliosus* E-value	Gene ontology	Hit sequence name	Species of origin	Accession IDs
0.790	4	101	145	2.00E-21	1.00E-34	P: lipid metabolic process; P: induction of apoptosis; F: endopeptidase inhibitor activity; F: calcium ion binding; P: negative regulation of angiogenesis; P: tissue remodelling; P: fibrinolysis; P: lipid transport; P: proteolysis; F: apolipoprotein binding; P: myoblast differentiation; P: muscle maintenance; F: serine-type endopeptidase activity; C: extracellular region; F: plasmin activity; P: blood circulation; P: tissue regeneration	Plasminogen	*Danio rerio*	tr|Q8AVB0|Q8AVB0_DANRE
0.714	5	147	162	3.00E-35	2.00E-39	P: response to organic nitrogen; P: multidrug transport; F: xenobiotic-transporting ATPase activity; P: canalicular bile acid transport; F: ATP binding; F: bile acid-exporting ATPase activity; P: response to organic cyclic substance; P: response to estradiol stimulus; C: basolateral plasma membrane; P: response to lipopolysaccharide; C: integral to plasma membrane	Canalicular multispecific organic anion transporter 1	*Gasterosteus aculeatus*	ENSGACP00000009848

Table 2 *Continued*

K_a/K_s	Number of SNPs	*Amphilophus astorquii* Bit score	*Amphilophus zaliosus* Bit score	*Amphilophus astorquii* E-value	*Amphilophus zaliosus* E-value	Gene ontology	Hit sequence name	Species of origin	Accession IDs
0.702	4	119	122	1.00E-26	1.00E-27	F: actin monomer binding; P: cardiac muscle contraction; P: regulation of the force of heart contraction; C: myosin complex; P: regulation of striated muscle contraction; F: calcium ion binding; P: ventricular cardiac muscle morphogenesis; F: motor activity; C: A band; C: I band	Myosin light chain 4	*Oryzias latipes*	ENSORLP00000021991
0.668	4	142	195	1.00E-33	1.00E-49	C: proteasome core complex; P: ubiquitin-dependent protein catabolic process; C: nucleus; C: cytosol; F: threonine endopeptidase activity	Proteasome subunit beta type-6 Precursor	*Takifugu rubripes*	ENSTRUP00000003687
0.632	3	98.2	139	2.00E-20	8.00E-33	Not available	A-kinase anchor protein 13	*Gasterosteus aculeatus*	ENSGACP00000000813
0.621	5	179	162	7.00E-45	1.00E-39	C: proteasome core complex; P: ubiquitin-dependent protein catabolic process; C: nucleus; C: cytosol; F: threonine endopeptidase activity; P: immune response	Proteasome subunit beta type-9 precursor	*Gasterosteus aculeatus*	ENSGACP00000000192
0.598	2	143	158	5.00E-34	2.00E-38	Not available	Novel protein-coding stickleback	*Gasterosteus aculeatus*	ENSGACP00000024881
0.548	3	116	105	1.00E-25	2.00E-22	F: molecular_function; C: cellular_component	Coiled-coil domain containing 93	*Gasterosteus aculeatus*	ENSGACP00000001928

EST, expressed sequence tags; SNP, single nucleotide polymorphism.

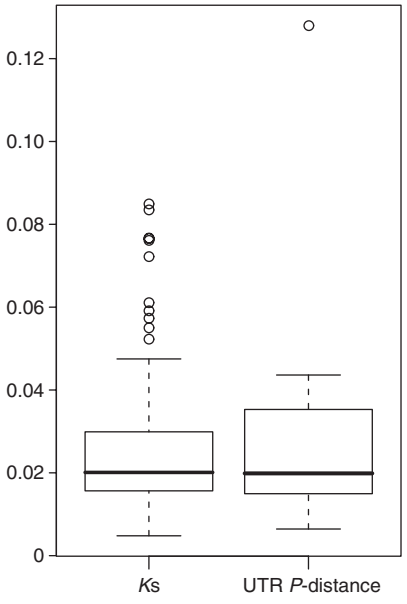

Fig. 4 Mean (±1 SD) K_s value and corrected p-distance of 3′-untranslated regions for orthologous ESTs. The means of these mutation rates are the same, suggesting that they both indicate neutral and consistent substitution.

indicating that *A. zaliosus* is only weakly diverged at neutral loci from the other Midas cichlids in Lake Apoyo (Barluenga *et al.* 2006). Nonetheless, this transcriptome-wide approach has provided the first indication of putatively functional genome divergences between these sympatric species.

Transcriptome sequencing and annotation of ESTs for a nonmodel organism

Pyrosequencing of our normalized cDNA library resulted in 24 174 ESTs for *A. astorquii* and 21 382 ESTs for *A. zaliosus*. We found that 13 106 pairs of ESTs were putatively orthologous between species. This represents a significant increase in the number of available cichlid ESTs (generated primarily by a few traditional EST studies using Sanger sequencing technology for East African cichlids, e.g. Renn *et al.* 2004; Watanabe *et al.* 2004; Salzburger *et al.* 2008) and inferred from chimeric genomic sequences from the cichlid genome project (Loh *et al.* 2008). Currently, approximately 45 000 ESTs are available for three African cichlid species, *Astatotilapia burtoni*, *Haplochromis chilotes* and *Haplochromis* 'redtail sheller' (http://compbio.dfci.harvard.edu/tgi/tgipage.html and NCBI dbEST; accessed 9 June 2009). Thus, our study contributes both a greater number of new cichlid ESTs to the research domain and, to our knowledge, the first neotropical cichlid ESTs.

Because of its relatively long read lengths, 454 pyrosequencing is the best method for *de novo* assembly (Rokas & Abbot 2009), although transcriptome assembly is nonetheless difficult and requires deep coverage (Weber *et al.* 2007; Wheat in press). Given that there is a constant upper limit (approximately 100 Mb for GS FLX standard chemistry) to the total number of base pairs generated per run, we sequenced normalized cDNA to try and maximize coverage of transcripts and reduce sequences of abundant transcripts. There will inevitably be a trade-off in cDNA preparation methods: whether or not to normalize, and the type of normalization approach to use. Largely, this depends on the experimental question being pursued, the diversity of input material and the number of sequence reads to be returned (Hale *et al.* 2009; Wheat in press). In this study, we used a 3′ extension approach, which purports to maximize depth and overall number of base pairs although at a cost of ORF length relative to random priming approaches. The equal coverage between species, the comparably high number of total base pairs, and the 3′-UTR bias in our results are because of our choice of library preparation. Nonetheless, more sequencing to gain deeper coverage and greater assembly power will be required to generate full-length ESTs for these cichlid species. Additionally, without high-coverage genomic sequences, we cannot confirm that our orthologous ESTs are in fact derived from the orthologous locus (i.e. expressed from the identical location in the genome) in both species. In the near future, with bacterial artificial chromosome (BAC) resources (K. Stölting, F. Henning, M. Lang, S. Fukamachi, A. Meyer, in preparation) and an emerging cichlid genome as a reference sequence, we will be able to more confidently identify genes important in the divergence of these two species and aim to map their functional importance.

Natural selection and the transcriptome

Of the 44 EST orthologues for which K_a and K_s could be calculated, six have $K_a/K_s > 1$, suggestive of positive selection acting on those genes by elevating the number of nonsynonymous substitutions. Ours may be a very conservative estimate because of the strict criteria we used for included orthologues. Eight candidate loci have K_a/K_s between 0.5 and 1. We included ESTs with $K_a/K_s > 0.5$ in our candidate genes of interest because we lack full-length genes, which will artefactually decrease K_a/K_s and cause one to overlook genic information relevant for positive selection (Swanson *et al.* 2004). The ESTs we identified as being under strong positive selection ($K_a/K_s > 1$) function in biosynthetic and metabolic processes, cognition, response to hormone stimuli and in the nervous system. Eight ESTs with K_a/K_s between 0.5 and 1 function in metabolic processes, tissue, blood and hormone regulation and

regeneration. These genes under natural selection will be of particular interest for future research (Jensen *et al.* 2007), although molecular laboratory approaches such as RACE PCR will probably be needed to get full-length sequences. Putatively, Darwinian selection or adaptive molecular evolution has resulted in important sequence differences between species in these genomic regions (Hurst 2009) and/or these regions are relevant to speciation (Noor & Feder 2006).

The ratio of nonsynonymous to synonymous substitutions is considered to be a good indicator of selective pressure at the sequence level (Yang & Bielawski 2000; Bustamante *et al.* 2005) and has been used to identify protein-coding genes under positive and purifying selection in a breadth of organisms (Hurst 2009). Evolutionary factors may limit the ability to detect signals of selection on a gene. For example, selection may act on putatively silent sites, selection pressures along a gene or gene fragment may be heterogeneous, and adaptive evolution may be limited to few functional sites (Yang & Bielawski 2000; Hurst 2002; Ellegren 2008). Additionally, evolutionarily important changes may lie in the gene regulatory region rather than the protein-coding region itself (Prud'homme *et al.* 2007). Further, gene sequence variation between species only demonstrates selection that has occurred in the past; to detect on-going or recent selection, population genetic data and comparisons are needed (Nielsen *et al.* 2007). These population- vs. species-level differences will be a focus of our future research.

Genome-wide mutation rate estimate of neotropical and African cichlids

We used our interspecific distance estimates based on neutral substitution (K_s and UTR) to calculate a transcriptome-wide estimate of substitution rate in these Midas cichlids. Thus, the minimum substitution rate would be 0.63×10^{-6} per site per year when calibrated to the maximal age of crater lake Apoyo (*c.* 20 000 years; Kutterolf *et al.* 2007). Based on a more biologically probable speciation time between *A. zaliosus* and *Amphilophus 'citrinellus'* (which includes all benthic morphs in the lake: *A. astorquii*, *Amphilophus chancho* and *Amphilophus flaveolus*) of 10 000 years (Barluenga *et al.* 2006), the substitution rate would be 1.25×10^{-6} per site per year. If the true divergence time is less than that estimated from mitochondrial DNA then the substitution rate would be faster. To our knowledge, this is the first transcriptome-wide inference of substitution rate for cichlid fishes.

The rate we inferred between Midas sympatric cichlid species based on substitution in protein-coding genes is considerably faster than the few previously published genome-wide estimates available across taxa (e.g. mammals generally: $\sim 2.2 \times 10^{-9}$ per site per year (Kumar & Subramanian 2002); humans specifically: $\sim 3.0 \times 10^{-8}$ per site per generation (Xue *et al.* 2009). Little consensus or knowledge exists regarding true genome- or transcriptome-level molecular clocks in vertebrates (Kumar 2005; Pulquerio & Nichols 2007). Our rate may be elevated by population-level polymorphism and therefore fall into the much debated possible discrepancy between population genetic and phylogenetic mutation rate estimates (e.g. Ho *et al.* 2005; Bandelt 2008) that has evidence in fish mitochondrial DNA (e.g. Burridge *et al.* 2008). Regardless, the substitution rates per year that we have generated will be useful to researchers of adaptive radiations of fish in general and cichlids in particular, for which interspecific divergences are characteristically shallow.

Although our estimates may be elevated by population polymorphism, the molecular evolutionary rate of fish is known to be fast: protein sequences in fish evolve significantly faster than their orthologues in mammals, both for duplicated genes and those retained in single copy (Ravi & Venkatesh 2008). For example, protein sequences between the two pufferfish species with sequenced genomes (*Takifugu rubripes* and *Tetraodon nigroviridis*) are more divergent than their homologues in mammals, although the pufferfishes diverged 32 million years ago whereas the mammals (human and mouse) diverged 61 million years ago (reviewed in Ravi & Venkatesh 2008).

Studying adaptive radiations with massively parallel sequencing

Based on the phenotypic diversity of cichlids, many have pondered whether there are some unusual properties in cichlid molecular evolution, genome size or plasticity (e.g. lineage-specific genome duplications, karyotypic dynamism, accelerated mutation rate) that allows for such spectacular diversification: yet all information to date indicates there is not (Kuraku & Meyer 2008). Studies seeking to understand adaptive sequence evolution as a source of variation to explain the diversity of cichlids have met with mixed success. Some cichlid studies have found signs of positive selection in genes involved in ecologically important traits, for example, MHC and fertilization (Gerrard & Meyer 2007), egg-dummy colour (Salzburger *et al.* 2007), colour perception genes (Terai *et al.* 2002a; Terai *et al.* 2006; Seehausen *et al.* 2008; Spady *et al.* 2005) and jaw bone development (Terai *et al.* 2002b). Others have found a lack of species-specific sequence differentiation (Watanabe *et al.* 2004; Loh *et al.* 2008; Kobayashi *et al.* 2009), which suggests that genomic or transcriptomic factors other than protein-coding sequence variation are responsible for extant cichlid diversity, such as expression levels (Kijimoto *et al.* 2005; Kobayashi *et al.* 2006)

or regulatory variation (Terai *et al.* 2003; Sanetra *et al.* 2009). Using a single pyrosequencing experiment, we have identified more genes putatively under natural selection than any of the above studies.

Genomic tools such as pyrosequencing offer a breadth of new, exciting avenues for research into the genetic basis underlying this great range of cichlid diversity, both in sequence polymorphisms and gene expression variation. By using genome- or transcriptome-wide techniques, we can now identify a greater number of candidate genes more quickly, with fewer biases, and at less cost than ever before (Fontanillas *et al.* 2010, Wheat in press). We can also have better success at discerning molecular differences between very closely related species, such as these recently diverged cichlid fish. Coupled with studies of parallel evolution, functional effects and fitness experiments, we are embarking on a new era of understanding how natural selection on the genome drives speciation. Hopefully, by integrating ecological speciation theory (Schluter 2000; Nosil *et al.* 2009) and emerging genomic and transcriptomic resources, we will begin to understand how genetic and ecological factors interact and might drive speciation.

Acknowledgements

This research was funded by a University of Konstanz Young Scholar's Award and a Natural Sciences and Engineering Research Council fellowship to KRE, University of Konstanz Zukunftskolleg fellowships to HMG and JCJ, and grants of the Deutsche Forschungsgemeinschaft to AM. We thank T. Lehtonen, M. Barluenga, and W. Salzburger for photographs used in Fig. 1 and three anonymous reviewers for suggestions on the manuscript. AM thanks the Institute for Advanced Study Berlin for a fellowship that supported him during this study and Wissenshaftskollge fellows J. Feder, J. Mallet and P. Nosil for fruitful discussions.

Conflicts of interest

The authors have no conflict of interest to declare and note that the sponsors of the issue had no role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References

Altenhoff AM, Dessimoz C (2009) Phylogenetic and functional assessment of orthologs inference projects and methods. *PLoS Computational Biology*, **5**, e1000262.

Altschul SF, Madden TL, Schaffer AA *et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, **25**, 3389–3402.

Bandelt H-J (2008) Time dependency of molecular rate estimates: tempest in a teacup. *Heredity*, **100**, 1–2.

Barlow GW, Munsey JW (1976) The red devil-Midas-arrow cichlid species complex in Nicaragua. In: *Investigations of the Ichthyology of Nicaraguan Lakes* (ed. Thorson TB), pp. 359–369. University of Nebraska Press, Lincoln.

Barluenga M, Meyer A (2004) The Midas cichlid species complex: incipient speciation in Nicaraguan cichlid fishes? *Molecular Ecology*, **13**, 2061–2076.

Barluenga M, Stölting KN, Salzburger W, Muschick M, Meyer A (2006) Sympatric speciation in Nicaragua crater lake cichlid fish. *Nature*, **439**, 719–723.

Barnes WM (1994) PCR amplification of up to 35-kb DNA with high fidelity and high yield from λ bacteriophage templates. *Proceedings of the National Academy of Sciences of the United States of America*, **91**, 2216–2220.

Baylis JR (1976) A quantitative study of long-term courtship: I. Ethological isolation between sympatric populations of the Midas cichlid, *Cichlasoma citrinellum*, and the arrow cichlid. *Behaviour*, **59**, 59–69.

Bouck A, Vision T (2007) The molecular ecologist's guide to expressed sequence tags. *Molecular Ecology*, **16**, 907–924.

Bunje PME, Barluenga M, Meyer A (2007) Sampling genetic diversity in the sympatrically and allopatrically speciating Midas cichlid species complex over a 16 year time series. *BMC Evolutionary Biology*, **7**, 25.

Burridge CP, Craw D, Fletcher D, Waters JM (2008) Geological dates and molecular rates: fish DNA sheds light on time dependency. *Molecular Biology and Evolution*, **25**, 624–633.

Bustamante CD, Fledel-Alon A, Williamson S *et al.* (2005) Natural selection on protein-coding genes in the human genome. *Nature*, **437**, 1153–1157.

Cerda J, Mercade J, Lozano J *et al.* (2008) Genomic resources for a commercial flatfish, the Senegalese sole (*Solea senegalensis*): EST sequencing, oligo microarray design, and development of the Soleamold bioinformatic platform. *BMC Genomics*, **9**, 508.

Chen FC, Li WH (2001) Genomic divergences between humans and other hominoids and the effective population size of the common ancestor of humans and chimpanzees. *American Journal of Human Genetics*, **68**, 444–456.

Chevreux B, Pfisterer T, Drescher B *et al.* (2004) Using the miraEST Assembler for Reliable and Automated mRNA Transcript Assembly and SNP Detection in Sequenced ESTs. *Genome Research*, **14**, 1147–1159.

Clabaut C, Bunje PME, Salzburger W, Meyer A (2007) Geometric morphometric analyses provide evidence for the adaptive character of the Tanganyikan cichlid fish radiations. *Evolution*, **61**, 560–578.

Collins FS, Green ED, Guttmacher AE, Guyer MS (2003) A vision for the future of genomics research. *Nature*, **422**, 835–847.

Conesa A, Götz S, García-Gómez JM, Terol J, Talón M, Robles M (2005) Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics*, **21**, 3674–3676.

Ellegren H (2008) Comparative genomics and the study of evolution by natural selection. *Molecular Ecology*, **17**, 4586–4596.

Elmer KR, Kusche H, Lehtonen TK, Meyer A (in press) Local variation and parallel evolution: morphological and genetic diversity across a species complex of neotropical crater lake cichlid fishes. *Philosophical Transactions of the Royal Society of London Series B*.

Fontanillas P, Landry CP, Witthopp PJ *et al.* (2010) Key considerations for measuring allelic expression on a genomic scale using high-throughput sequencing. *Molecular Ecology*, **19** (Suppl. 1), 212–227.

Gerrard DT, Meyer A (2007) Positive selection and gene conversion in SPP120, a fertilization-related gene, during the east African cichlid fish radiation. *Molecular Biology and Evolution*, **24**, 2286–2297.

Götz S, García-Gómez JM, Terol J *et al.* (2008) High-throughput functional annotation and data mining with the Blast2GO suite. *Nucleic Acids Research*, **36**, 3420–3435.

Hale MC, McCormick CR, Jackson JR, De Woody JA (2009) Next generation pyrosequencing of gonad transcriptomes in the polyploid lake sturgeon (*Acipenser fulvenscens*): the relative merits of normalization and rarefaction in gene discovery. *BMC Genomics*, **10**, 203.

Hellmann I, Zöllner S, Enard W, Ebersberger I, Nickel B, Pääbo S (2003) Selection on human genes as revealed by comparisons to chimpanzee cDNA. *Genome Research*, **7**, 831–837.

Ho SYW, Phillips MJ, Cooper A, Drummond AJ (2005) Time dependency of molecular rate estimates and systematic overestimation of recent divergence times. *Molecular Biology and Evolution*, **22**, 1561–1568.

Hubbard T, Andrews D, Caccamo M *et al.* (2005) Ensembl 2005. *Nucleic Acids Research*, **33**, D447–D453.

Hudson ME (2007) Sequencing breakthroughs for genomic ecology and evolutionary biology. *Molecular Ecology Notes*, **8**, 3–17.

Hurst LD (2002) The Ka/Ks ratio: diagnosing the form of sequence evolution. *Trends in Genetics*, **18**, 486–487.

Hurst LD (2009) Evolutionary genomics and the reach of selection. *Journal of Biology*, **8**, 12.

International Cichlid Genome Consortium (2006) Genetic basis of vertebrate diversity: The Cichlid Fish Model (March 28, 2006). 24 pp. Available from http://www.genome.gov/Pages/Research/Sequencing/SeqProposals/CichlidGenomeSeq.pdf.

Jensen JD, Wong A, Aquadro CF (2007) Approaches for identifying targets of positive selection. *Trends in Ecology & Evolution*, **23**, 568–577.

Jukes TH, Cantor CR (1969) Evolution of protein molecules. In: *Mammalian Protein Metabolism, III* (ed. Munro HN), pp. 21–132. Academy Press, New York.

Kijimoto T, Watanabe M, Fujimura K *et al.* (2005) cimp1, a novel astacin family metalloproteinase gene from east African cichlids, is differentially expressed between species during growth. *Molecular Biology and Evolution*, **22**, 1649–1660.

Kobayashi N, Watanabe M, Kijimoto T *et al.* (2006) *magp4* gene may contribute to the diversification of cichlid morphs and their speciation. *Gene*, **373**, 126–133.

Kobayashi N, Watanabe M, Horiike T, Kohara Y, Okada N (2009) Extensive analysis of EST sequences reveals that all cichlid species in Lake Victoria share almost identical transcript sets. *Gene*, **441**, 187–191.

Kumar S (2005) Molecular clocks: four decades of evolution. *Nature Reviews Genetics*, **6**, 654–662.

Kumar S, Subramanian S (2002) Mutation rates in mammalian genomes. *Proceedings of the National Academy of Sciences of the United States of America*, **99**, 803–808.

Künster A, Wolf JBW, Backström N *et al.* (2010) Comparative genomics based on massive parallel transcriptome sequencing reveals patterns of substitution and selection across 10 bind species. *Molecular Ecology*, **19** (Suppl. 1), 266–276.

Kuraku S, Meyer A (2008) Genomic analysis of cichlid fish 'natural mutants'. *Current Opinion in Genetics & Development*, **18**, 551–558.

Kutterolf K, Freundt A, Pérez W, Wehrmann H, Schmincke H-U (2007) Late Pleistocene to Holocene temporal succession and magnitudes of highly-explosive volcanic eruptions in west-central Nicaragua. *Journal of Volcanology and Geothermal Research*, **163**, 55–82.

Larkin MA, Blackshields G, Brown NP *et al.* (2007) Clustal W and Clustal X version 2.0. *Bioinformatics*, **23**, 2947–2948.

Loh Y-HE, Katz LS, Mims MC, TD K, Yi SV, Streelman JT (2008) Comparative analysis reveals signatures of differentiation amid genomic polymorphism in Lake Malawi cichlids. *Genome Biology*, **9**, R113.

Lynch M, Sung W, Morris K *et al.* (2008) A genome-wide view of the spectrum of spontaneous mutations in yeast. *Proceedings of the National Academy of Sciences of the United States of America*, **105**, 9272.

McCrary JK, López L (2008) *El Monitoreo De Las Mojarras (Amphilophus Spp.) En Nicaragua Con Aportees Sobre Su Ecología Y Estado De Conservación En La Laguna De Apoyo*. Ministerio del Ambiente y los Recursos Naturales (MARENA), Managua, Nicaragua, pp. 43–50.

Moore MJ, Dhingra A, Soltis PS *et al.* (2006) Rapid and accurate pyrosequencing of angiosperm plastid genomes. *BMC Plant Biology*, **6**, 17.

Moore MJ, Bell CD, Soltis PS, Soltis DE (2007) Using plastid genome-scale data to resolve enigmatic relationships among basal angiosperms. *Proceedings of the National Academy of Sciences of the United States of America*, **104**, 19363–19368.

Nielsen R, Hellmann I, Hubisz M, Bustamante C, Clark AG (2007) Recent and ongoing selection in the human genome. *Nature Reviews Genetics*, **8**, 857–868.

Noor MAF, Feder JL (2006) Speciation genetics: evolving approaches. *Nature Reviews Genetics*, **7**, 851–861.

Nosil P, Harmon LJ, Seehausen O (2009) Ecological explanations for (incomplete) speciation. *Trends in Ecology & Evolution*, **24**, 145–156.

Oldfield RG (2009) Captive breeding observations support the validity of a recently described cichlid species in Lake Apoyo, Nicaragua. *Occasional Papers of the Museum of Zoology University of Michigan*, **741**, 1–14.

Parsons KJ, Robinson BW, Hrbek T (2003) Getting into shape: an empirical comparison of traditional truss-based morphometric methods with a newer geometric method applied to New World cichlids. *Environmental Biology of Fishes*, **67**, 417–431.

Prud'homme B, Gompel N, Carroll SB (2007) Emerging principles of regulatory evolution. *Proceedings of the National Academy of Sciences of the United States of America*, **104**, 8605–8612.

Pulquerio MJF, Nichols RA (2007) Dates from the molecular clock: how wrong can we be? *Trends in Ecology & Evolution*, **22**, 180–184.

Ravi V, Venkatesh B (2008) Rapidly evolving fish genomes and teleost diversity. *Current Opinion in Genetics & Development*, **18**, 544–550.

Renn SCP, Aubin-Horth N, Hofmann HA (2004) Biologically meaningful expression profiling across species using heterologous hybridization to a cDNA microarray. *BMC Genomics*, **5**, 42.

Rokas A, Abbot P (2009) Harnessing genomics for evolutionary insights. *Trends in Ecology & Evolution*, **24**, 192–200.

Rüber L, Verheyen E, Meyer A (1999) Replicated evolution of trophic specializations in an endemic cichlid fish lineage from Lake Tanganyika. *Proceedings of the National Academy of Sciences of the United States of America*, **96**, 10230–10235.

Salzburger W (2009) The interaction of sexually and naturally selected traits in the adaptive radiations of cichlid fishes. *Molecular Ecology*, **18**, 169–185.

Salzburger W, Braasch I, Meyer A (2007) Adaptive sequence evolution in a color gene involved in the formation of the characteristic egg-dummies of male haplochromine cichlid fishes. *BMC Biology*, **5**, 51.

Salzburger W, Renn SCP, Steinke D, Braasch I, Hofmann HA, Meyer A (2008) Annotation of expressed sequence tags for the East African cichlid fish *Astatotilapia burtoni* and evolutionary analyses of cichlid ORFs. *BMC Genomics*, **9**, 96.

Sanetra M, Henning F, Fukamachi S, Meyer A (2009) A microsatellite-based genetic linkage map of the cichlid fish, *Astatotilapia burtoni* (Teleostei): a comparison of genomic architectures among rapidly speciating cichlids. *Genetics*, **182**, 387–397.

Schluter D (2000) *The Ecology of Adaptive Radiations*. Oxford University Press Inc., New York.

Seehausen O, Terai Y, Magalhaes IS *et al.* (2008) Speciation through sensory drive in cichlid fish. *Nature*, **455**, 620–626.

Shendure J, Porreca GJ, Reppas NB *et al.* (2005) Accurate multiplex colony sequencing of an evolved bacterial genome. *Science*, **309**, 1728–1732.

Shin H, Hirst M, Bainbridge M, Magrini V, Mardis E (2008) Transcriptome analysis for *Caenorhabditis elegans* based on novel expressed sequence tags (ESTs). *BMC Biology*, **6**, 30.

Smit AFA, Hubley R, Green P (1996) RepeatMasker Open-3.0. Available from http://www.repeatmasker.org.

Spady TC, Seehausen O, Loew ER, Jordan RC, Kocher TD, Carleton KL (2005) Adaptive molecular evolution in the opsin genes of rapidly speciating cichlid species. *Molecular Biology and Evolution*, **22**, 1412–1422.

Stauffer Jr JR, McKaye KR (2002) Descriptions of three new species of cichlid fishes (Teleostei: Cichlidae) from lake Xiloá, Nicaragua. *Dirección de Investigación de la Universidad Centroamericana, Managua, Nicaragua*, **12**, 1–18.

Stauffer Jr JR, McCrary JK, Black KE (2008) Three new species of cichlid fish (Teleostei: Cichlidae) in Lake Apoyo, Nicaragua. *Proceedings of the Biological Society of Washington*, **121**, 117–129.

Swanson WJ, Wong A, Wolfner MF, Aquadro CF (2004) Evolutionary expressed sequence tag analysis of *Drosophila* female reproductive tracts identifies genes subjected to positive selection. *Genetics*, **168**, 1457–1465.

Terai Y, Mayer WE, Klein J, Tichy H, Okada N (2002a) The effect of selection on a long wavelength-sensitive (LWS) opsin gene of Lake Victoria cichlid fishes. *Proceedings of the National Academy of Sciences of the United States of America*, **99**, 15501–15506.

Terai Y, Morikawa N, Okada N (2002b) The evolution of the pro-domain of bone morphogenetic protein 4 (Bmp4) in an explosively speciated lineage of East African cichlid fishes. *Molecular Biology and Evolution*, **19**, 1628–1632.

Terai Y, Morikawa N, Kawakami K, Okada N (2003) The complexity of alternative splicing of hagoromo mRNAs is increased in an explosively speciated lineage in East African cichlids. *Proceedings of the National Academy of Sciences of the United States of America*, **100**, 12798–12803.

Terai Y, Sasaki T, Takahashi K *et al.* (2006) Divergent selection on opsins drives incipient speciation in Lake Victoria cichlids. *PLoS Biology*, **4**, 2244–2251.

The Gene Ontology Consortium (2000) Gene ontology: tool for the unification of biology. *Nature Genetics*, **25**, 25–29.

The UniProt Consortium (2008) The Universal Protein Resource (UniProt). *Nucleic Acids Research*, **36**, D190–D195.

Toth AL, Varala K, Newman TC *et al.* (2007) Wasp gene expression supports an evolutionary link between maternal behavior and eusociality. *Science*, **318**, 441–444.

Vera JC, Wheat CW, Fescemyer HW *et al.* (2008) Rapid transcriptome characterization for a nonmodel organism using 454 pyrosequencing. *Molecular Ecology*, **17**, 1636–1647.

Vivas R, McKaye KR (2001) Habitat selection, feeding ecology, and fry survivorship in the *Amphilophus citrinellus* species complex in Lake Xiloá. *Journal of Aquaculture and Aquatic Sciences*, **IX**, 32–48.

Watanabe M, Kobayashi N, Shin-i T *et al.* (2004) Extensive analysis of ORF sequences from two different cichlid species in Lake Victoria provides molecular evidence for a recent radiation event of the Victoria species flock: identity of EST sequences between *Haplochromis chilotes* and *Haplochromis* sp. ''Redtailsheller''. *Gene*, **343**, 263–269.

Weber APM, Weber KL, Carr K, Wilkerson C, Ohlrogge JB (2007) Sampling the *Arabidopsis* transcriptome with massively parallel pyrosequencing. *Plant Physiology*, **144**, 32–42.

Wheat CW (in press) Rapidly developing functional genomics in ecological model systems via 454 transcriptome sequencing. *Genetica*; DOI 10.1007/s10709-008-9326-y.

Wilson AB, Noack-Kunnmann K, Meyer A (2000) Incipient speciation in sympatric Nicaraguan crater lake cichlid fishes: sexual selection versus ecological diversification. *Proceedings of the Royal Society of London. Series B*, **267**, 2133–2141.

Wolf JBW, Bayer T, Haubold B *et al.* (2010) Nucleotide divergence versus gene expression differentiation: comparative transcriptome sequencing in natural isolates from the carrion crow and its hybrid zone with the hooded crow. *Molecular Ecology*, **19** (Suppl. 1), 162–175.

Xue Y, Wang Q, Long Q *et al.* (2009) Human Y chromosome base-substitution mutation rate measured by direct sequencing in a deep-rooting pedigree. *Current Biology*, **19**, 1453–1457.

Yang Z (2007) PAML 4: phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution*, **24**, 1586–1591.

Yang Z, Bielawski JP (2000) Statistical methods for detecting molecular adaptation. *Trends in Ecology & Evolution*, **15**, 496–503.

Yang Z, Nielsen R (2000) Estimating synonymous and non-synonymous substitution rates under realistic evolutionary models. *Molecular Biology and Evolution*, **17**, 32–43.

The Meyer Laboratory (AM, KRE, JCJ, HGM, SF) is interested in the molecular basis of the vast ecological adaptations and evolutionary radiations of cichlid fishes, both African and neotropical. The Kuraku Laboratory (SK, SB) studies the evolution of gene repertoires, especially in early vertebrates, and the integration of bioinformatics and molecular evolutionary developmental biology.
