First publ. in: Molecular Biology and Evolution 24 (2007), 10, pp. 2286–2297

Positive Selection and Gene Conversion in SPP120, a Fertilization-Related Gene, during the East African Cichlid Fish Radiation

Dave T. Gerrard and Axel Meyer

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

The ability to infer historical natural selection from sequence data aides in finding genes that might be important in adaptation and the formation of new species. As the fastest evolving and largest known vertebrate radiation, the cichlid fish of the African Great Lakes exhibit a wide range of recent morphological diversification. We used DNA databases, mostly of expressed sequence tags, to find candidate orthologous coding sequences from 2 tribes of cichlids and, using an automated procedure, scanned these sequence pairs for high dN/dS, the signal of positive selection and protein adaptation. The results included vertebrate genes commonly found to be under selection (e.g., major histocompatibility complex [MHC] loci) as well as genes known to be important specifically in the cichlid radiation (e.g., long-wave-sensitive opsins). Further investigation focused on a gene encoding a fertilization-related protein, SPP120, which was previously known only from cichlids. Using maximum likelihood analysis on novel SPP120 cDNA sequences from a range of African cichlids, we demonstrate the influence of positive selection in a specific subregion of the protein. We also show that SPP120 is a tandemly arranged, multicopy gene evolving with occasional interlocus gene conversion. A search of the Medaka genome database also revealed a tandem arrangement of multiple SPP120 copies and evolutionary rate differences between Medaka gene subregions mirroring those found for cichlids. Combined, these results suggest that SPP120 has been under repeated diversifying selection for over 100 Myr.

Introduction

It is a major goal of evolutionary genetics to match genetic differences between species or individuals with those phenotypic differences that, when seen by natural selection, drive adaptive evolution (Schluter 2001; Colosimo et al. 2005). The cichlid fish species flocks of the East African Great Lakes underwent a series of independent explosive radiations unsurpassed throughout vertebrate history in magnitude and rate of speciation (Fryer and Iles 1972; Meyer 1993; Kocher 2004; Salzburger and Meyer 2004; Seehausen 2006). In several African lakes, these adaptive radiations of hundreds of species each show repeated coevolution of their feeding morphology, coloration, and behavior in parallel (Meyer et al. 1990; Kocher et al. 1993; Meyer 1993; Stiassny and Meyer 1999). These adaptive features have been suspected to have been shaped by ecological and/or sexual selection (Stiassny and Meyer 1999; Kocher 2004). The combination of the diversity of forms within lakes and the convergence of forms between lakes has long been considered a great opportunity to test the importance of natural selection on speciation (Schluter 2000). However, the specification of any particular divergent trait can involve complex gene pathways, and no agreement has been reached as to whether any one family or class of genes are overrepresented in the formation of genetic barriers delineating young species or whether speciation is largely due to an accumulation of random differences (Coyne and Orr 1998, 2004; Swanson and Vacquier 2002). Genes involved in fertilization have been shown to be of particular importance during the evolution of reproductive barriers (Swanson and Vacquier 2002; Ting et al. 2004) and, hence, might also play a role in speciation of cichlid fishes.

In model organisms, it is possible to investigate candidate genes for certain species-specific phenotypes. However, due to constraints inherent to nonmodel organisms, this type of approach is often more difficult to apply. Traits that are potentially causally related to the diversification of cichlids include genes that are involved in variation of jaw morphology (Albertson et al. 2003) and visual sensitivities to different wavelengths of light (Spady et al. 2005). Linkage analysis has also been successfully employed to find genomic loci influencing species differences (Albertson et al. 2003; Streelman et al. 2003), but the relatively long lifecycle and small brood size of most of the African cichlids (Fryer and Iles 1972), along with the low resolution of current genetic maps, make this approach tedious (Lee et al. 2005).

Alternatively, the expanding and diversifying sequence databases present an opportunity to scan large number of genes and compare their rates of evolution (Swanson et al. 2001). In an effort to identify novel genes of potential importance in the adaptive radiation of cichlids, we compared published coding sequences of Haplochromine and Tilapiine cichlids, taking advantage of 3 recently created expressed sequence tag (EST) libraries from Haplochromine species (Renn et al. 2004; Watanabe et al. 2004). After identifying probable orthologues, we automated the pairwise alignment of these sequences and calculated the ratio of the rates of substitution at nonsynonymous and synonymous sites (dN/dS) to identify genes evolving under positive Darwinian selection. Here, we present the results of this approach and the investigation of a reproductive protein, SPP120, that showed a very high dN/dS ratio but that was also found to be a multicopy gene evolving under the influence of gene conversion.

Materials and Methods

Scan for High dN/dS Candidate Genes

DNA sequences originating from 2 tribes of African cichlids were downloaded from GenBank. Not all available sequences from the Haplochromini, the extremely species-rich lineage of cichlid that makes up most of the adaptive radiations of cichlids in East African Great Lakes (Salzburger et al. 2005), were used because, often, only a single locus

Key words: reproductive genes, natural selection, gene conversion, cichlid, SPP120.

E-mail: axel.meyer@uni-konstanz.de.

Table 1
Major Sources of GenBank DNA Sequences for Haplochromini and Tilapiini Cichlid Tribes

Quantity	Species	Type/Source	Citation
Haplochromini group (39,468)			
21,653	*Haplochromis chilotes*	Jaw ESTs (fry up to 60 days)	Watanabe et al. 2004
14,073	*Haplochromis sp. 'redtail sheller'*	Jaw ESTs (fry up to 60 days)	Watanabe et al. 2004
3,670	*Astatotilapia burtoni*	Brain ESTs (all ages)	Renn et al. 2004
Tilapiini group (4,665)			
294	*Oreochromis niloticus*	Brain ESTs (adult)	Hamilton et al. 2000
1,377	*O. niloticus*	Genomic	Patent WO03060160 (863 $\leq$ 200 bp)
1,241	*O. niloticus*	Genomic/BAC shotgun sequencing	Lee 2004
323	*O. niloticus*	Genomic sequence tag site	Lee et al. 2005

has been sequenced in dozens of similar species/individuals as part of a population study. In autumn 2005, over 90% of the Haplochromini sequences in GenBank came from the EST libraries of 3 species (Renn et al. 2004; Watanabe et al. 2004). Therefore, all nucleotide entries (including non-EST sequences) from *Astatotilapia burtoni*, *Ptyochromis (Haplochromis) sp. 'redtail sheller'*, and *Paralabidochromis (Haplochromis) chilotes* totaling 39,468 sequences at that time were downloaded. All the 4,665 available Tilapiini sequences were downloaded and these came from a variety of sources, such as genomic surveys, sequence tag site markers, and patents (*Oreochromis niloticus*, the Nile Tilapia is a commercially important food fish); over 95% of sequences came from *O. niloticus* or *Oreochromis mossambicus*. The Tilapiini are an older lineage of cichlids and are not particularly species rich, at least by comparison to the Haplochromini. The major sources of these sequences are listed in table 1.

After a screen for cloning vector sequences, approximately 400 of the *Ptyochromis (Haplochromis) sp. 'redtail sheller'* sequences (Watanabe et al. 2004) had to be removed from the data sets because they were almost entirely composed of sequencing vector. Two Blast databases were constructed using software available from the National Center for Biotechnology Information, one database for the haplochromine and one for the tilapiine sequences. The shorter list of Tilapiini sequences was then used for Blast searches of all these sequences against the Haplochromini database only retaining matches with an *e* value $\leq 10^{-15}$. The highest scoring significant Haplochromini sequence matching each query was then added to a list of sequences (992 in total), which were Blast searched back against the Tilapiini database. This produced a list of pairs of sequences that could be filtered to identify the reciprocal best matches from the Blast searches. Nonreciprocal or duplicate matches were removed. Additionally, all the Tilapiini sequences from the list of matches were Blast searched back against the Tilapiini database to find all matches that were higher scoring than the match found in the Haplochromini database (controlling for sequence length). These matches were sorted into families of highly similar sequences, and only the longest matching sequence was kept in the list to control for redundant sequencing, misannotation, and gene families where one sequence from one database was a top hit for many sequences from the other database. All sequences matching the full *O. mossambicus* mitochondrial genome (AY597335) were also removed from the data set (1,201 sequences).

The following sequence manipulations were carried out using a Perl script with the aid of the BioPerl modules (Stajich et al. 2002). The sequences were aligned according to a forced amino acid translation of the longest open reading frames. If the coding sequence was annotated, it was used; otherwise, the sequences were aligned in 6 coding arrangements and the highest scoring protein alignment was used to align the nucleotide sequences (using needle and tranalign, EMBOSS—Rice et al. 2000). Under automation, pairwise dN/dS was calculated using the YN00 method of Yang and Nielsen (2000) implemented in PAML (Yang 1997). Because of the low number of orthologous pairs found, those with dS less than 0.2 and more than 200 bp of aligned coding sequence were realigned manually and dN/dS was recalculated in MEGA according to the "Modified Nei–Gojobori" method (Nei and Gojobori 1986).

Target Gene SPP120

The pairwise dN/dS is a weak estimate of positive selection because most nucleotide sites in most functional genes are under strong purifying selection (Makalowski and Boguski 1998). When dN/dS is averaged over all sites, any signal derived from the minority of adaptive substitutions is diluted by the pervasive signal of constraint (Li 1997). Greater resolution can be achieved by analyzing the accumulation of changes to a protein across a phylogenetic tree. The codeml program, part of the PAML package (Yang 1997), uses a maximum likelihood approach to contrast the ability of alternative, nested models of evolution in explaining the available data. This method is able to discern the signal of adaptation from only some sites against a background of constraint and is a more powerful test for the presence of positive selection (Yang and Nielsen 2002).

We chose a gene, SPP120, for further study. SPP120 is expressed in the gonad, was previously only known from Tilapia, and was experimentally shown to encode a protein with sperm-binding affinity (Mochida et al. 1999). As such, it is a good candidate for positive selection (Swanson and Vacquier 2002). The maximum likelihood method requires, however, a divergent sampling of sequences from species within the framework of a strongly resolved phylogeny. Therefore, we amplified SPP120 from different African cichlid lineages using the phylogeny of Salzburger et al. (2005). The relationships of cichlids within Lakes Malawi and Victoria are weakly resolved, but, because the cichlid radiations of Lakes Victoria and Malawi are monophyletic

Table 2
Primers used to Amplify SPP120 from Cichlids

Name	Sequence (5′-3′)	Notes
01F	ACA ATC ACC AGA GAG CAT CCT GAG C	Based on the 5′ untranslated region of SPP120 cDNA from *Oreochromis niloticus*
97F	ATG CTC TGC CCA CTG CTC TTC C	Includes the ATG start codon
2520R	AGT CAG TCC ATC CAA GTC ACC ATG G	Includes the terminator codon
2580R	AAG CAG CAT GTC TCT TGT CTC TGC	Based on the 3′ untranslated region of SPP120 cDNA from *O. niloticus*
420F	AGA ACA GTT TGA GAT GCA GTG TGC	In exon 4, based on *O. niloticus* exon/intron structure elucidated in this study
1207R	TGG AGT TTC CAC ACA GAC CCT CC	In exon 10, based on *O. niloticus* exon/intron structure elucidated in this study
1000F	CTG GAC GAC TCC ACA CTG ACC	Highly conserved internal primer. Used for *Pseudocrenilabrus multicolor* amplification
2470R	CAG GAC ACG CTG ATG AAG GC	Highly conserved internal primer. Used for *P. multicolor* amplification

NOTE.—F and R refers to forward and reverse primers, respectively; the number corresponds to the approximate position along the published *O. niloticus* SPP120 coding sequence (AB073751). The sequences and positions of sequencing primers are available from the authors on request.

(Meyer et al. 1990; Moran and Kornfield 1993; Sultmann et al. 1995; Verheyen et al. 2003; Salzburger et al. 2005), 2 species from each lake could be sampled as nearest neighbors. Novel SPP120 sequences were obtained from the following East African species: *A. burtoni* (a widespread nonendemic species), *Pundamilia nyererei* (Lake Victoria), *Haplochromis sp. '44'* (Lake Victoria), *Melanochromis auratus* (Lake Malawi), *Pseudotropheus sp. 'Bicolor'* (Lake Malawi), and *Pseudocrenilabrus multicolor* (a widespread nonendemic species).

A single adult male of each species was anaesthetized with approximately 0.04% Tricaine (Sigma, Frankfurt, Germany) and then placed into ice-cold water for 5–10 min. The spinal column was severed before dissection. The paired testes were located running anterior–posterior from immediately beneath the swim bladder to near the proctodeum. RNA was extracted from one homogenized testis using Trizol reagent (Invitrogen, Karlsruhe, Germany) and stored at −80 °C. cDNA was reverse transcribed from the testis RNA using Superscript II Reverse Transcriptase (Invitrogen). SPP120 was polymerase chain reaction (PCR) amplified from the cDNA using primers designed on the single, full-length, published sequence from Tilapia (AB073751—Mochida et al. 2002). PCR products over 2 kb in length were cloned using the TOPO-TA Kit (Invitrogen), and several clones were checked by sequencing with M13 forward and reverse primers. SPP120-positive clones were partially sequenced using primers designed on the *O. niloticus* sequence and then completed by incremental PCR steps with additional primers. Sequences from all species were compiled into one GAP4 (Staden Package—Staden et al. 2000) database to simultaneously check for sequencing gaps or errors and to make multispecies alignments. All clones were fully sequenced in both directions. Table 2 lists the primers used to PCR amplify SPP120 from testis cDNA or from genomic DNA.

Genomic Sequences

SPP120 was amplified from *O. niloticus* genomic DNA using the Fermentas Long PCR Enzyme Mix with an extension time of 11 min for the first 10 cycles, then an additional 5 s for each of the remaining 25 cycles. The PCR products were cloned and sequenced. The raw sequences from the genomic clones were compiled into a GAP4 database along with the completed cDNA sequences to identify the exon/intron boundaries and so that new exonic conserved primers could be designed.

The lengths of the genes were variable and generally longer in haplochromine species than in *O. niloticus*. Therefore, a portion of the gene from exon 4 to exon 10 (primers 420F/1207R, table 2) was PCR amplified and cloned from all species. This region corresponds to amino acids 107–370 and contained 1,422–1,981 bp of intronic sequence. All novel sequences have been deposited in GenBank under accession numbers EF486251–EF486264 (cDNAs) and EF490572–EF490597 (genomic DNAs).

Screen of the Astatotilapia Bacterial Artificial Chromosome Library

To better estimate the number of copies of SPP120 in the cichlid genome, we conducted a hybridization screen of the recently developed bacterial artificial chromosome (BAC) library of *A. burtoni* following the method of Lang et al. (2006). As multicopy genes may include retrotransposed (spliced) copies of themselves, the screen was conducted using a mixture of genomic (intron containing) and cDNA (intronless)-derived probes spanning exons 7–12. Positive clones were screened by PCR targeting the same region from exon 4 to exon 10 (primers 420F/1207R, table 2).

Sequence Analyses

The full-length cDNA sequences were extracted from the GAP4 database ready aligned and then analyzed using MEGA3 (Kumar et al. 2004). The coding frame and protein sequence from the published SPP120 sequence were used to check for frameshift and nonsense mutations in the sequences. Phylogenetic trees of the coding regions from the full-length cDNA sequences were constructed using the Neighbor-Joining and Minimum-Evolution methods implemented in MEGA3 using the Tamura–Nei distance (Tamura and Nei 1993).

Table 3
The 10 Loci with Highest dN/dS Computed after Automated Alignment

Locus	gi: Tilapiini/Haplochromini	Automated dN/dS[a]	Manual dN/dS[b]	Aligned Length	Notes
MHC class IIB locus 1	3282880/46534539	1.304	2.481	246	Poorly aligned segments removed in manual alignment
SPP120	17298175/46540369	1.120	0.860[c]	596	Gonad expressed (Mochida et al. 2002)
Unknown	74059555/46525962	1.114	2.665	327	Ensembl family ENSF00000000187. Similar to TRIM protein family
Betaglobin	30142117/46509156	1.103	0.545	436	Wrong coding frame 1 bp longer than the correct one
Pan-epithelial glycoprotein	62079565/46536586	1.075	1.428	546	Pan-epithelial glycoprotein/ Tumor-associated calcium signal transducer ovary derived
Type I collagen alpha	3201257/46525611	0.975	—	—	3′ untranslated region
MHC class IIB locus 4	3282888/46537432	0.966	3.203	249	Poorly aligned segments removed in manual alignment
Aromatase P450 Type 1	4838537/4185561	0.812	1.328	717	Ovarian form
NCCRP 1 like	23096105/46533282	0.782	0.901	642	Immune response (Ishimoto et al. 2004)
LWS opsin	7542659/50841447	0.765	0.779	1071	Recently under positive selection (Spady et al. 2005)

[a] YN00, Yang and Nielsen 2000.
[b] Modified Nei–Gojobori (implemented in MEGA3).
[c] A sequencing stutter at the most 5′ of the Haplochromine sequence alters the coding frame for a short region.

Recombination

The presence of recombination among the cDNA sequences was suggested by an inspection of the variable sites along the sequence alignment. The program "permute" by McVean et al. (2002) was used to test for a significant correlation of linkage disequilibrium with physical distance along the alignment, a clear signal of recombination.

Selection

The relative abilities of different models to explain the evolution of the cDNA sequences were contrasted using the program codeml (PAML; Yang 1997). First, models M0 and M3 were compared with test for heterogeneity between codon sites in the value of the dN/dS ratio, ω. Second, a mixed model allowing for some sites under positive selection ($\omega > 1.0$) and other sites under varying degrees of purifying selection ($\omega < 1.0$) or neutrality ($\omega \sim 1.0$) Model M8 was contrasted with a similar model with the same flexibility in ω below 1.0 but with the constraint that the ω at fast-evolving sites could not be greater than 1.0 (Model M8a, see Wong et al. 2004). This is effectively a test between neutrality and positive selection at fast-evolving sites against a background of constrained sites in both models (Wong et al. 2004). The Neighbor-Joining tree constructed from these sequences was used as input to codeml (see Results).

Medaka Database Scan

The October 2006 release (#41) of the Medaka (*Oryzias latipes*) genome (http://www.ensembl.org/Oryzias_latipes; Kasahara et al. 2007) was Blast searched using the protein sequence of the *O. niloticus* (Tilapia) SPP120 sequence. Significant hits to the protein sequence were assigned to particular exons using the exon/intron structure that we determined for *O. niloticus* (see above).

Results

The database Blast searching and filtering produced 353 pairs of putative Haplochromine/Tilapiine orthologues of varying divergence and alignment quality. Of these, 82 sequence pairs produced uninterrupted coding sequence alignments of sufficient length to calculate dN/dS. The median dS value was 6.3%. Among the top 10 pairs (ranked by dN/dS) were 3 genes that were previously shown to be under selection in their coding regions during the radiation of cichlids indicating that this approach is able to identify positively selected genes (table 3, and supplementary table 1, Supplementary Material online). Two were major histocompatibility complex (MHC) Class II loci (Figueroa et al. 2000) and a third, the long-wave-sensitive (LWS) opsin, was one of the opsin genes previously found to be evolving under the influence of positive selection in cichlids (Terai, Mayer, et al. 2002; Spady et al. 2005). Another gene previously indicated to be under selection during the cichlid radiation, bone morphogenic protein 4 (Terai, Morikawa Okada 2002), generated a pairwise dN/dS of just 0.113. This perhaps reflects well the inability of the pairwise method to detect small adaptive changes against a very strong background of constraint. Similarly, the "hagoromo" gene, which has been linked to striped color patterns and contains a short subregion undergoing fast evolution (Terai, Morikawa, et al. 2002), did not stand out in our list of candidates for selection ($dN/dS = 0.268$).

The low number of putative orthologues meant that all 82 could be checked manually before further investigation of strong candidates of positive selection. The results of these manual alignments are shown as a plot of dN against dS in figure 1. Though the initial scan revealed several good candidates for further study, SPP120 was chosen because of the length of the alignment, the provenance of the tilapia cDNA sequence, and the functional studies linking this gene to cichlid reproduction (Mochida et al. 1999, 2002).

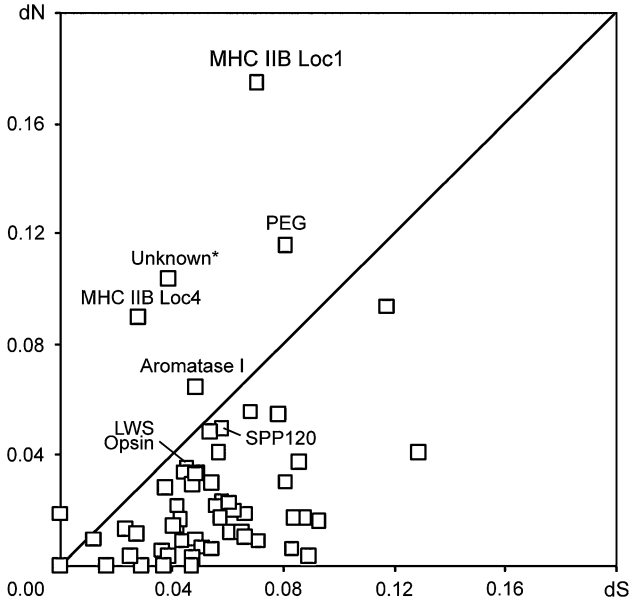

FIG. 1.—Scatterplot of dN against dS for the 58 manual alignments sharing more than 200 bp and with dS less than 0.2. The mean dS is 0.054, and the mean dN is 0.028. The diagonal line represents equality and a dN/dS of 1.0. Loci discussed in the text or appearing in table 3 are annotated. Asterisk denotes that the gene labeled "unknown" is a member of the tripartite motif-containing (TRIM) protein family (position 3 in table 3 and supplementary table 1, Supplementary Material online).

Full-Length cDNA Sequences and gene Trees

PCR amplification of SPP120 from cichlid testis cDNA revealed several unexpected properties of this gene. First, multiple, divergent transcribed copies of the gene were amplified from almost all cDNA samples. Three sequences from *M. auratus* and 2 from *P. bicolor* were full-length transcribed sequences that differed by more than 1% at the nucleotide level (fig. 2). We failed to amplify SPP120 from testis cDNA of *P. multicolor* but were able to amplify several distinct copies of the gene from its genomic DNA. Two divergent copies of this gene in this species included 95% of the coding sequence, conserved without nonsense mutations or frameshifts. Among the Lake Victoria haplochromine cichlids (including *A. burtoni*), most cDNA sequences differed by no more than 1% either between or within species, and it is difficult to distinguish distinct loci from alleles based on this data alone. The numbers of distinct sequences (those with divergences >0.3% within species) were 4, 2 and, 2 for *A. burtoni*, *P. nyererei*, and *Haplochromis sp. '44'*, respectively.

Inspection of the sequences revealed 2 independent mutations that likely represent a transition to a pseudogene of their respective sequences (marked with ψ in fig. 2). cDNA copy number 2 from *Pseudotropheus sp. 'Bicolor'* contains a 1-bp frameshifting deletion at position 468. The cDNA-derived sequences from *A. burtoni*, *P. nyererei*, and *Haplochromis sp. '44'* were generally very similar. cDNA copy 4 from *A. burtoni* has a G–A transition, creating a premature TGA stop codon at position 2131 (losing 11% of the coding sequence, including half the "Zona Pellucida" [ZP] domain). The same substitution is also seen in individual clones from both of the Lake Victoria species (*P. nyererei* cDNA copy 2 and *Haplochromis sp. '44'* copy 2), but, interestingly, a single cDNA cloned from *P. nyererei* (copy 1 in fig. 2) featured a second, potentially compensatory substitution in the same codon that alters the sequence to CGA (Arginine).

In addition, the sequences of *M. auratus* appear to be partly recombinant. This was confirmed by a significant negative correlation between linkage disequilibrium (r^2) and distance along the sequence (-0.21, $P < 0.002$). It appears that copy 3 is chimaeric, sharing greatest similarity to copy 2 within the first (most 5′ 1,800 bp) but is closer to copy 1 in the final 600 bp. All of the "pseudogenes" listed above, the recombinant sequence 3, and any sequences sharing more than 99.5% identity were excluded from the maximum likelihood analysis. As the relationships among the Lake Victoria Haplochromine sequences were ambiguous, only 2 sequences were used, *A. burtoni* cDNA copy 3 and the more divergent *Haplochromis sp. '44'* cDNA copy 1.

The gene tree topology, estimated from cDNA and *P. multicolor* genomic (see below) SPP120 sequences (fig. 2*A*), was used as input for the maximum likelihood analysis because it better represents the evolutionary relationships of the gene copies than does the species phylogeny (fig. 2*B*; Salzburger et al. 2005). Maximum likelihood analysis of the remaining 8 sequences revealed strong heterogeneity in ω among the codons and favored a model of evolution including an excess of amino acid changing substitutions at some sites, consistent with the influence and signature of adaptive evolution.

The maximum likelihood results are summarized in table 4. First, the heterogeneity test (M3 vs. M0) revealed that ω is not uniform along the coding sequence of SPP120 ($P < 10^{-5}$). Second, when a mixed model of evolution including sites with $\omega > 1$ (Model 8) was tested against the same model but with the upper class of ω constrained to 1.0 (Model 8a, reflecting neutral evolution at those sites), then the former had a significantly greater likelihood ($P < 10^{-5}$) using a more conservative test with 2 degrees of freedom (Wong et al. 2004). In both tests, the more parameter rich model was successful and gave very similar estimates of ω for the fast-evolving sites. Under Model 8, the value of ω predicted for the sites with adaptive substitutions is 5.92. The Bayes Empirical Bayes posterior probabilities (BEBpp—Yang et al. 2005) of sites with $\omega > 1.0$ list 17 codon positions with BEBpp > 0.95, of which 13 are within the first third of the protein, the region with no detectable homology to other known proteins. The positions of codons predicted to be under the influence of strong positive selection are illustrated in figure 3.

Genomic Sequence from *O. niloticus*

A genomic PCR product containing the entire SPP120-coding region was amplified from Tilapia DNA and overlapping sequences from *P. multicolor* spanning most of the gene. In Tilapia, primers SPP120_01F and SPP120_2520R were used, generating a 9.3-kb product, which, aligned to the published cDNA sequence, revealed the presence of 18 introns (fig. 4). Multiple copies were

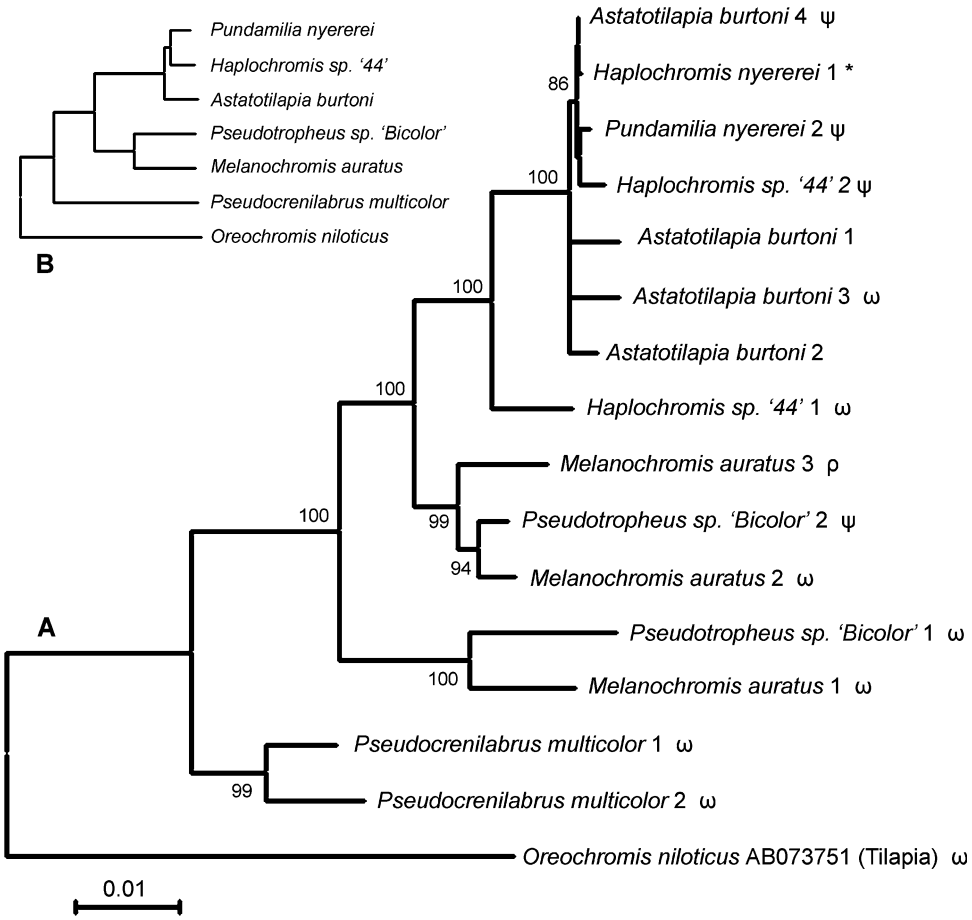

FIG. 2.—(*A*) Neighbor-Joining phylogeny (2,307 bp, Tamura–Nei distance) of sequences obtained by PCR amplifying SPP120 from testis-derived cDNA of male African cichlids (scale: substitutions per site). Only bootstrap supports greater than 70% (of 1,000) are shown at the nodes. Only same-species sequences with greater than 0.3% divergence (Tamura–Nei) are shown, that is, more than 7 differences. All sequences from a species come from a single male individual. Those marked "ω" were used in the maximum likelihood analyses. Those marked "ψ" are potential transcribed pseudogene copies—see text. "ρ" denotes a recombinant sequence sharing similarity to sequences 1 and 2 from *Melanochromis auratus*. Asterisk denotes that cDNA sequence *Pundamilia nyererei* 1 features a "compensatory" substitution altering the premature stop codon to "CGA." (*B*) The known species phylogeny is based on Salzburger et al. 2005.

only as distinct as one might expect from alleles (0.24% between 5,000 bp sequenced in 2 different clones, Tamura–Nei distance). The coding sequence of the fully sequenced clone differed from the published sequence (Mochida et al. 2002) by 14 amino acids. There are 14 nonsynonymous and 4 synonymous differences in the coding region. When scaled per site, this gives a dN/dS ratio of 1.6. Previously, Mochida K (personal communication) detected a second copy of the SPP120 protein within *O. niloticus*, but only one cDNA was sequenced. In *P. multicolor*, we were unable to amplify the entire gene in a single PCR. Instead, exons 2–10 were amplified using primers 97F and 1207R (table 2). Exons 8–18 were amplified using primers 1000F and 2470R (table 2). For each segment, 2 or more distinct clones were obtained. Two distinct pairs of 5′ and 3′ clones showed 100% identity in the overlapping exons 8–10 but divergence between pairs. These were used to generate 2 sequences spanning exons 2–18 and were used to generate the potential coding sequences used in figure 2 and in the maximum likelihood analysis.

Amplification of a portion of SPP120 from genomic DNA of the same male individuals produced a sequence tree (fig. 5) with identical or highly similar genomic counterparts to most of the testis-derived sequences (except *M. auratus* cDNA copy 2). In addition, several novel and distinct (>1% divergence) copies were amplified from several

Table 4
Summary of Maximum Likelihood Analyses on SPP120 cDNA Sequences

Test	First Model	Second Model	$2\Delta\ell$ (P value)	Omega Distribution
Heterogeneity	M0	M3	115.056 ($P < 0.0001$)	85.5% sites: $\omega = 0.49$ and 14.5% sites: $\omega = 5.92$
Selection versus neutrality	M8a	M8	72.096 ($P < 0.0001$)	85.5% sites: $0.43 < \omega < 0.55$ and 14.5% sites: $\omega = 5.92$

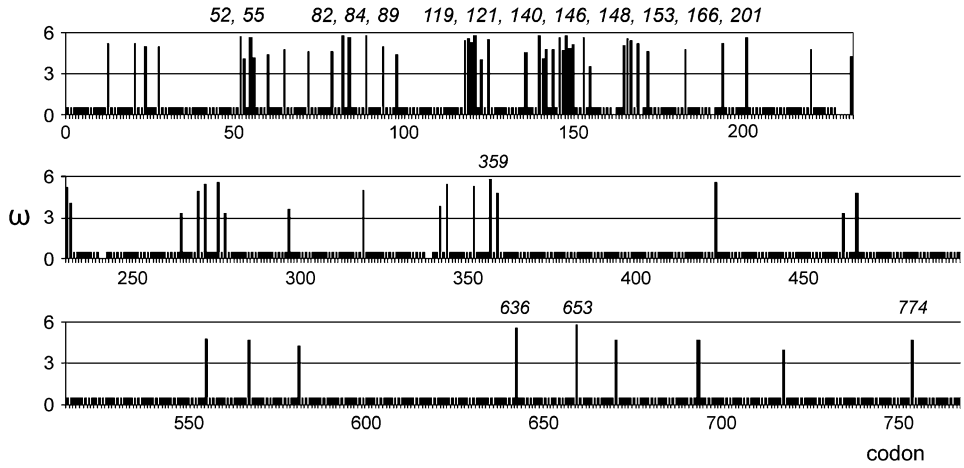

FIG. 3.—SPP120 codon sites predicted to be under positive selection ($\omega > 1.0$) during the divergence of East African cichlids. The second and third rows correspond to the VWD and ZP domains, respectively. A total of 17 sites with BEBpp > 95% that $\omega > 1.0$ are numbered (italics) according to the published protein sequence of Mochida et al. (2002). All but 4 of these occur within the N-terminal uncharacterized domain.

species. Taking together, both cDNA-derived and genomic DNA–derived sequences (all from one male per species) brings the copy number for the exon 4 to exon 10 region to 5 for *M. auratus*, 4 for *P. multicolor*, and 3 each for *P. bicolor*, *P. nyererei*, and *Haplochromis sp. '44'*.

BAC Library Results

The screen of the *A. burtoni* BAC library produced 8 clones positive for the mixed spliced/nonspliced SPP120 probe. Initial analysis by PCR revealed length and copy number differences between SPP120 forms on different clones. Seemingly full-length versions of SPP120 were found on 3 clones (149-C9, 164-G7, and 182-L14). Long range PCR from these BAC clones gave single products of over 20 kb. PCR and sequencing of the subregion from exon 4 to exon 10, revealed that all 3 clones carried the same copy of SPP120 and that the coding region most closely matched the *A. burtoni* cDNA copy 4 (fig. 2). The other clones contained "spliced" forms of SPP120 lacking introns, curtailed forms containing only introns 8–10 with a long interspersed nuclear element (LINE) element in the place of exons 5–7, or both of these forms. Two BAC clones (183-P24 and 165-O8) contained both the spliced and the truncated forms of SPP120. Sequencing of the BAC PCR products revealed that copies were divergent both between and within different clones. BAC clone 182-O1 contained a spliced form matching most closely cDNA copy 1 from *Haplochromis sp. '44'*. Spliced forms from BAC clones 183-P24 and 165-O8 were similar to each other and to the alternative genomic copies from *P. nyererei* and *Haplochromis sp. '44'* (numbered genomic copy II in fig. 5). The nonspliced but curtailed forms of SPP120, however, did not share discriminatory synapomorphies with any of the previously sequenced Lake Victoria forms and were less similar to other *A. burtoni* sequences than the closest of those from the Lake Malawi species. These fragments of SPP120 may therefore be relics dating back to shortly before or during the diversification of the modern haplochromines.

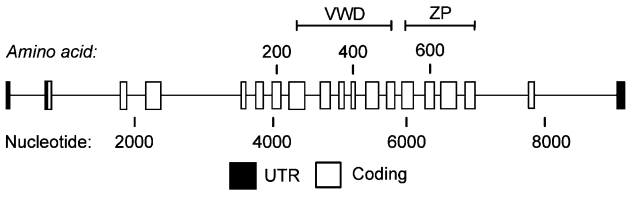

FIG. 4.—Exon structure of *Oreochromis niloticus* SPP120 deduced by alignment of novel genomic sequence to the published cDNA, AB073751 (drawn using Gene Structure Draw by Vamsi Veeramachaneni, http://warta.bio.psu.edu/cgi-bin/Tools/StrDraw.pl). UTR, Untranslated region.

Database Searches

When this project began, there were no obvious non-cichlid orthologous sequences matching SPP120 in any genome database (Mochida et al. 2002). However, Blast searches using protein sequence revealed strong hits to SPP120 in the October 2006 release of the Medaka genome expanding the taxonomic range of SPP120 to the ancestor of these lineages. In the Medaka genome, there are multiple strong matches (Blast P value $< 10^{-80}$ in most cases) to SPP120 exons, and these are all located within a 5-Mb region on chromosome 18. Other significant Blast hits (P value $< 10^{-05}$) are divergent paralogues matching the von Willebrand factor D (VWD) or ZP domains. There are multiple potential copies of the SPP120 gene in the chromosome 18 region, each with a similar level of divergence from the cichlid sequence. The relative positions and orientations of clusters of exon hits are shown in figure 6 and a Neighbor-Joining phylogeny of the more complete clusters in figure 7. It is notable that the clusters located in the same orientation along the chromosome ([1,2] or [5,7,8,10]) show greater similarity than with clusters in the opposite direction. The outgroup sequences shown in figure 7, from *Gasterosteus aculeatus* and *Danio rerio*, are not full-length copies of SPP120 and only correspond to the most 3′ 1100 nt of the coding region (mostly the ZP domain). More complete copies of the gene, if they exist, have yet to be found.

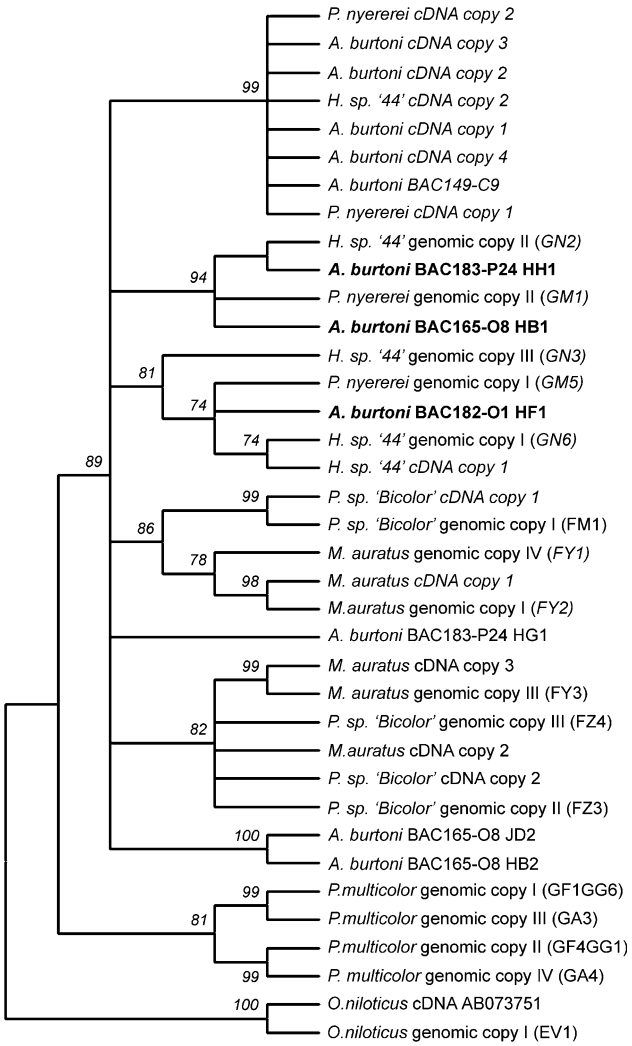

FIG. 5.—Condensed Neighbor-Joining phylogeny featuring sequences obtained by PCR amplifying exons 4–10 of SPP120 from genomic DNA of African cichlids (with original clone names), the cDNA sequences from figure 2 and sequences obtained by PCR from the SPP120-positive *Astatotilapia burtoni* BAC clones. The tree is based on 419 bp shared by all these classes of sequences. BAC clone subsequences that represent "spliced" genomic forms, lacking introns are indicated in bold. Only sequences differing by more than 0.5% (Tamura–Nei distance) are shown, and nodes with less than 70% bootstrap support are collapsed.

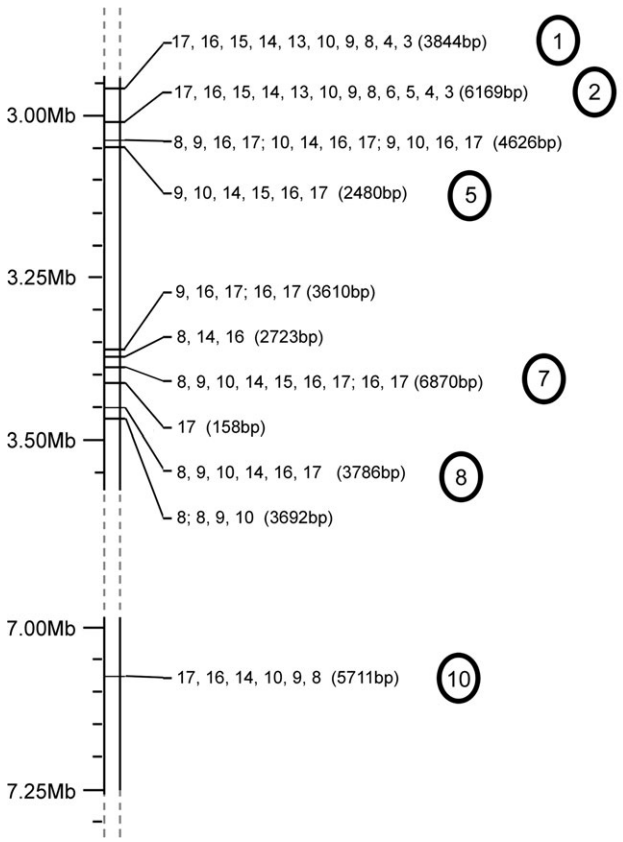

FIG. 6.—Relative positions of significant Blast hits on Medaka chromosome 18 generated by querying the *O. niloticus* SPP120 protein sequence against the Medaka genome (October 2006). The numbers represent matches to specific exons of SPP120 and are listed (left to right) in their chromosomal orientation (proximal to distal). Clusters (groups of matches shown here on the same line) are separated by 10 kb and/or a change of orientation of the Blast hits.

An alignment of the entire coding region of the cichlid SPP120 to 3 of the Medaka clusters (1, 2, and 8) contains numerous short, frame-preserving indels between the 4 sequences, the Kozak sequence (Kozak 1981) is intact (ncctccaa(a/c)ATG), and the AG/GT exon splice markers are preserved for almost all exons. For clusters 1 and 2, which are similar, there are 3 exon/intron boundaries that differ from the cichlid sequence, and in each case, there is an alternative splice site within 4 codons distance. Exon 18 of these clusters lacks an in-frame GT intron start signal, but the last codon of this exon is the translation terminator codon, which is preserved. Cluster 8 features 2 frameshift indels relative to the other sequences in exons 4 and 7 and is entirely lacking what would be intron 9. Based purely on this in silico comparison of sequence differences, clusters 1 and 2 could produce proteins in Medaka of almost identical lengths to SPP120 in cichlids.

Discussion

We aimed to identify genes that might be important in the adaptive radiation of East African cichlids by scanning for genes with an atypical rate of amino acid evolution. The discovery of several genes that had been previously identified to have been under recent natural selection by this method provided assurance that this method had sufficient power to do so. The other genes with high dN/dS values were mostly reproductive or immune system genes in line with expectations for this statistic (Yang and Bielawski 2000). As whole-genome EST libraries are often derived from brain or gonad tissue because of the wide-ranging gene expression of these tissues, this might increase the likelihood of finding reproductive genes. However, with the exception of LWS opsin and aromatase I, which both came from gene-targeting studies, all the haplochromine sequences listed in table 3 originated from EST libraries generated by Watanabe et al. (2004) from juvenile cichlid jaw tissue. This list is not presented as strong evidence of

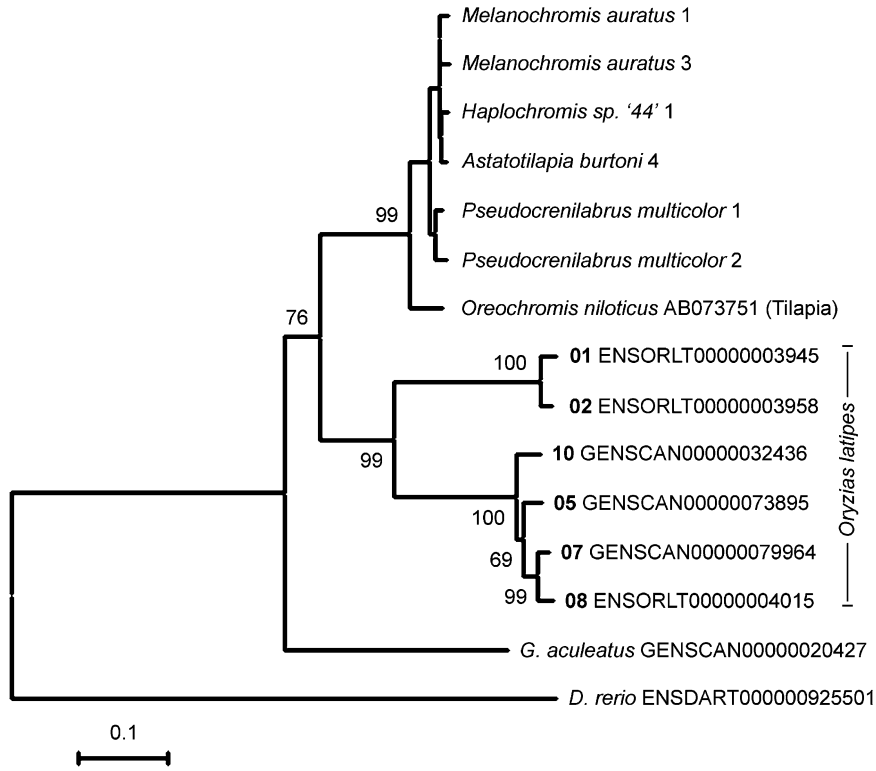

FIG. 7.—Neighbor-Joining tree based on 1,133 coding nucleotide sites shared by the different Medaka SPP120 clusters, the cichlid sequences from figure 2 and the best Blast homologues from the *Gasterosteus aculeatus* (Stickleback) and *Danio rerio* (Zebrafish) Ensembl transcript databases (scale: substitutions per site). The Medaka sequences are denoted by the cluster groups from figure 6 and then by the database accession numbers given to in silico predicted genes (Ensembl, October 2006).

adaptation of these genes in cichlids but solely as an indicator of a pattern worth investigation.

The molecular genetic evidence supporting widespread adaptive evolution of reproductive genes is matched only by that of genes involved in immune responses or their evasion. We chose to focus on a reproductive protein, SPP120, for further investigation because it was known to be functional and contained protein domains involved in binding to sperm and perhaps also to the ZP (or chorion) surrounding vertebrate oocytes (Mochida et al. 1999). The influence of positive Darwinian selection was confirmed by maximum likelihood analyses on novel cDNA sequences obtained from a variety of cichlids, but only after the influences of gene duplication, gene conversion and pseudogenization were accounted for.

The excess of nonsynonymous substitutions in this gene during the evolution of the cichlids is better represented by the mixed model of evolution that includes some sites (~15%) evolving under the influences of positive Darwinian selection. This signal remains one of the strongest yet observed in cichlids, even after a significant portion of the available data is discarded to remove any confounding influence of recombinant sequences.

From a structural perspective, the combination of the VWD domain and the ZP domain is a rare configuration, best known from alpha-tectorin, the major noncollagenous component of the tectorial membrane in mammalian ears (Legan et al. 1997). This gene has 4 VWD domains in tandem and a single C-terminus ZP domain. As alpha-tectorin seems to exist in teleosts (a predicted protein on Medaka Chr14 has 57% identity to human alpha-tectorin), SPP120 is not the orthologue of this gene but may be a paralog specific to higher teleosts. It has recently been speculated that the ZP domain was a feature of the ancestral animal oocyte coating due to its presence in mammalian, bird, and teleost oocyte proteins as well as those of the invertebrate abalone (Mold et al. 2001; Smith et al. 2005). The closest teleost homologues of the mammalian proteins have been described in Zebrafish (Mold et al. 2001) and in a percomorph (Modig et al. 2006), and these loci map to Medaka chromosomes 6 and 24 (by Blast homology). The existence of the ZP domain in SPP120 is therefore intriguing because of its strong expression in testis. Has this domain been co-opted due to its affinity for sperm or has it just filled a need for an extracellular structural glycoprotein as in other proteins unrelated to reproduction?

The precise function of the SPP120 "genes" remains unknown. At least one copy has the potential to produce a 790-amino acid protein in each of the cichlid species studied and does so successfully in *O. niloticus* (Mochida et al. 2002). Within cichlids, the VWD and ZP domains have evolved under purifying selection, but the 200–amino acid, N-terminus sequence has undergone adaptive diversifying evolution. A full-length coding sequence has also been conserved without stop codons or frameshift mutations for more than 100 Myr (Steinke et al. 2006) since the divergence of the cichlid and Medaka lineages (see below). Again, the divergence of SPP120, both between Medaka

and cichlids and between different loci within Medaka, is accelerated in the same 200-amino acid region. This pattern hints at coevolution between SPP120 and other genes, perhaps as a conflict over the ability for sperm to fertilize eggs. The findings of Mochida et al. (2002), that what was thought to be a single copy gene was expressed in both ovaries and gonads, might stand against a sex-specific role for SPP120. However, the evidence from Mochida et al. (2002) and our own observations, suggest that there could be quite a strong difference in expression between these 2 tissues.

Alternatively, the SPP120 genes may have altered their function or functions specifically during the cichlid radiation. Grier and Fishelson (1995) describe a distinction between mouth-brooding and substrate-spawning Tilapiine cichlids; the sperm of the former are packed, immotile, in a periodic acid-Schiff (PAS)-positive mucus (probably a glycoprotein) that is taken up by the female into her mouth where fertilization subsequently occurs. It might be speculated that 1 gene involved in this trait is SPP120. Though this particular trait is not seen outside Tilapiine cichlids (Wickler 1997), the fertilization and brooding strategies of African cichlids are diverse (Fryer and Iles 1972) and likely to have been under varied and perhaps repeated episodes of natural selection.

The evidence for occasional but repeated gene conversion during the evolution of this gene family is manifold. The recombinant cDNA sequences from a single male *M. auratus* individual suggested multiple divergent loci encoding different versions of SPP120, and this was confirmed for all species by genomic sequencing. The screen of the BAC library confirmed multiple loci, their tandem arrangement (for some at least) within 200 kb (the upper limit of *A. burtoni* BAC clone lengths), and the existence of spliced and retrotransposed gene copies. Finally, the recently released genome sequence of Medaka (Kasahara et al. 2007) showed that copies of SPP120 arranged in the same orientation along chromosome 18 exhibit greater similarity than copies lying in opposite orientations (fig. 6), even when the inverted copies are located much closer. This suggests that the Medaka SPP120 family underwent a history of gene conversion events dating back tens of millions of years, maintained after an inversion event separated clusters into forward- and reverse-orientated copies. Furthermore, it is possible that the homogenizing force of gene conversion has been operating on a tandem array of SPP120 duplicates since the most recent common ancestor of cichlids and Medaka dating back over 100 Myr (Steinke et al. 2006) and potentially of influence over more than 7,000 species of percomorph fish. The frequency of gene conversion events between SPP120 loci is, however, seemingly limited relative to the rate of cladogenesis (figs. 2 and 7).

Though the prevalence of gene conversion throughout the genome is hard to determine, it is intriguing that several genes and gene families that are thought to be under positive selection have also recently been shown to evolve concertedly. For example, the MHC Class II genes, a paradigm of coding region adaptation in vertebrates, show a high degree of allelic copy number variation in teleosts (Malaga-Trillo et al. 1998; Reusch et al. 2004), and it has been proposed that for some MHC loci in sticklebacks (*G. aculeatus*), gene conversion may be more important than mutation in generating diversity, the fuel of adaptive evolution (Reusch and Langefors 2005). Similar co-occurrences of diversifying selection and gene conversion have been observed in protocadherins (Noonan et al. 2004; Wu 2005) and the vitelline envelope forming genes of Abalone, another classic example of diversifying selection (Aagaard et al. 2006). In mammalian zonadhesin, a sperm surface protein that binds species specifically with the egg ZP, tandem VWD domain repeats have evolved both divergently and concertedly (Herlyn and Zischler 2006). It will be interesting to find out whether arrays of tandem gene copies have a greater propensity to undergo rapid, adaptive evolution or whether this correlation is just a by-product of the occasional neofunctionalization of duplicated gene copies (Stephens 1951; Ohno 1970).

Supplementary Material

Supplementary table 1 is available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

We thank Walter Salzburger for help in the choice of species and their dissection and Klaus Zanker for some of the laboratory work. The constructive criticism of 2 anonymous reviewers improved an earlier version of this manuscript. D.T.G. was supported by a Research Fellowship from the Alexander von Humboldt Stiftung and A.M. by grants from the Deutsche Forschungsgemeinschaft. Medaka data—"The data have been provided freely by the National Institute of Genetics and the University of Tokyo for use in this publication/correspondence only."

Literature Cited

Aagaard JE, Yi X, MacCoss MJ, Swanson WJ. 2006. Rapidly evolving zona pellucida domain proteins are a major component of the vitelline envelope of abalone eggs. Proc Natl Acad Sci USA. 103:17302–17307.

Albertson RC, Streelman JT, Kocher TD. 2003. Genetic basis of adaptive shape differences in the cichlid head. J Hered. 94:291–301.

Colosimo PF, Hosemann KE, Balabhadra S, Villarreal G Jr, Dickson M, Grimwood J, Schmutz J, Myers RM, Schluter D, Kingsley DM. 2005. Widespread parallel evolution in sticklebacks by repeated fixation of ectodysplasin alleles. Science. 307:1928–1933.

Coyne J, Orr A. 2004. Speciation. Sunderland (MA): Sinnauer Associates.

Coyne JA, Orr HA. 1998. The evolutionary genetics of speciation. Philos Trans R Soc Lond B Biol Sci. 353:287–305.

Figueroa F, Mayer WE, Sultmann H, O'HUigin C, Tichy H, Satta Y, Takezaki N, Takahata N, Klein J. 2000. Mhc class ii b gene evolution in East African cichlid fishes. Immunogenetics. 51:556–575.

Fryer G, Iles TD. 1972. The cichlid fishes of the great lakes of Africa: their biology and evolution. Edinburgh (UK): Oliver & Boyd.

Grier HJ, Fishelson L. 1995. Colloidal sperm-packaging in mouthbrooding tilapiine fishes. Copeia. 1995:966–970.

Hamilton LC, Macpherson GR, Wright JM. 2000. Expressed sequence tags derived from brain tissue of Oreochromis niloticus. J Fish Biol. 56:219–222.

Herlyn H, Zischler H. 2006. Tandem repetitive d domains of the sperm ligand zonadhesin evolve faster in the paralogue than in the orthologue comparison. J Mol Evol. 63:602–611.

Ishimoto Y, Savan R, Endo M, Sakai M. 2004. Non-specific cytotoxic cell receptor (nccrp)-1 type gene in tilapia *(Oreochromis niloticus)*: Its cloning and analysis. Fish Shellfish Immunol. 16:163–172.

Kasahara M, Naruse K, Sasaki S, et al. (38 co-authors). 2007. The medaka draft genome and insights into vertebrate genome evolution. Nature. 447:714–719.

Kocher TD. 2004. Adaptive evolution and explosive speciation: the cichlid fish model. Nat Rev Genet. 5:288–298.

Kocher TD, Conroy JA, McKaye KR, Stauffer JR. 1993. Similar morphologies of cichlid fish in lakes Tanganyika and Malawi are due to convergence. Mol Phylogenet Evol. 2:158–165.

Kozak M. 1981. Possible role of flanking nucleotides in recognition of the AUG initiator codon by eukaryotic ribosomes. Nucleic Acids Res. 9:5233–5252.

Kumar S, Tamura K, Nei M. 2004. MEGA3: integrated software for molecular evolutionary genetics analysis and sequence alignment. Brief Bioinform. 5:150–163.

Lang M, Miyake T, Braasch I, Tinnemore D, Siegel N, Salzburger W, Amemiya CT, Meyer A. 2006. A BAC library of the East African haplochromine cichlid fish Astatotilapia burtoni. J Exp Zoolog B Mol Dev Evol. 306:35–44.

Lee BY. 2004. Approach to the identification of sex-determining genes in the tilapia genome by genetic mapping and comparative positional cloning. Hubbard Center for Genome Studies. Durham (NH): University of New Hampshire.

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

Legan PK, Rau A, Keen JN, Richardson GP. 1997. The mouse tectorins. Modular matrix proteins of the inner ear homologous to components of the sperm-egg adhesion system. J Biol Chem. 272:8791–8801.

Li W-H. 1997. Molecular evolution. Sunderland (MA): Sinauer Associates.

Makalowski W, Boguski MS. 1998. Evolutionary parameters of the transcribed mammalian genome: an analysis of 2,820 orthologous rodent and human sequences. Proc Natl Acad Sci USA. 95:9407–9412.

Malaga-Trillo E, Zaleska-Rutczynska Z, McAndrew B, Vincek V, Figueroa F, Sultmann H, Klein J. 1998. Linkage relationships and haplotype polymorphism among cichlid Mhc class ii b loci. Genetics. 149:1527–1537.

McVean G, Awadalla P, Fearnhead P. 2002. A coalescent-based method for detecting and estimating recombination from gene sequences. Genetics. 160:1231–1241.

Meyer A. 1993. Phylogenetic relationships and evolutionary processes in East African cichlid fishes. Trends Ecol Evol. 8:279–284.

Meyer A, Kocher TD, Basasibwaki P, Wilson AC. 1990. Monophyletic origin of Lake Victoria cichlid fishes suggested by mitochondrial DNA sequences. Nature. 347: 550–553.

Mochida K, Kondo T, Matsubara T, Adachi S, Yamauchi K. 1999. A high molecular weight glycoprotein in seminal plasma is a sperm immobilizing factor in the teleost Nile tilapia, Oreochromis niloticus. Dev Growth Differ. 41:619–627.

Mochida K, Matsubara T, Andoh T, Ura K, Adachi S, Yamauchi K. 2002. A novel seminal plasma glycoprotein of a teleost, the Nile tilapia (Oreochromis niloticus), contains a partial von Willebrand factor type d domain and a zona pellucida-like domain. Mol Reprod Dev. 62:57–68.

Modig C, Modesto T, Canario A, Cerda J, Hofsten Jv, Olsson P-E. 2006. Molecular characterization and expression pattern of zona pellucida proteins in gilthead seabream (Sparus aurata). Biol Reprod. 75:717–725.

Mold DE, Kim IF, Tsai C-M, Lee D, Chang C-Y, Huang RCC. 2001. Cluster of genes encoding the major egg envelope protein of zebrafish. Mol Reprod Dev. 58:4–14.

Moran P, Kornfield I. 1993. Retention of an ancestral polymorphism in the mbuna species flock (Teleostei: Cichlidae) of Lake Malawi. Mol Biol Evol. 10:1015–1029.

Nei M, Gojobori T. 1986. Simple methods for estimating the number of synonymous and non-synonymous nucleotide substitutions. Mol Biol Evol. 3:418–426.

Noonan JP, Grimwood J, Schmutz J, Dickson M, Myers RM. 2004. Gene conversion and the evolution of protocadherin gene cluster diversity. Genome Res. 14:354–366.

Ohno S. 1970. Evolution by gene duplication. New York: Springer-Verlag.

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

Reusch TB, Langefors A. 2005. Inter- and intralocus recombination drive Mhc class iib gene diversification in a teleost, the three-spined stickleback Gasterosteus aculeatus. J Mol Evol. 61:531–541.

Reusch TB, Schaschl H, Wegner KM. 2004. Recent duplication and inter-locus gene conversion in major histocompatibility class ii genes in a teleost, the three-spined stickleback. Immunogenetics. 56:427–437.

Rice P, Longden I, Bleasby A. 2000. Emboss: the european molecular biology open software suite. Trends Genet. 16: 276–277.

Salzburger W, Mack T, Verheyen E, Meyer A. 2005. Out of Tanganyika: genesis, explosive speciation, key-innovations and phylogeography of the haplochromine cichlid fishes. BMC Evol Biol. 5:17.

Salzburger W, Meyer A. 2004. The species flocks of East African cichlid fishes: recent advances in molecular phylogenetics and population genetics. Naturwissenschaften. 91:277–290.

Schluter D. 2000. The ecology of adaptive radiation. Oxford: Oxford University Press.

Schluter D. 2001. Ecology and the origin of species. Trends Ecol Evol. 16:372–380.

Seehausen O. 2006. African cichlid fish: a model system in adaptive radiation research. Proc Biol Sci. 273:1987–1998.

Smith J, Paton IR, Hughes DC, Burt DW. 2005. Isolation and mapping the chicken zona pellucida genes: an insight into the evolution of orthologous genes in different species. Mol Reprod Dev. 70:133–145.

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. Mol Biol Evol. 22:1412–1422.

Staden R, Beal KF, Bonfield JK. 2000. The staden package, 1998. Methods Mol Biol. 132:115–130.

Stajich JE, Block D, Boulez K, et al. (21 co-authors). 2002. The bioperl toolkit: Perl modules for the life sciences. Genome Res. 12:1611–1618.

Steinke D, Salzburger W, Meyer A. 2006. Novel relationships among ten fish model species revealed based on a phylogenomic analysis using ESTs. J Mol Evol. 62:772–784.

Stephens SG. 1951. Possible significances of duplication in evolution. Adv Genet. 4:247–265.

Stiassny MLJ, Meyer A. 1999. Cichlids of the rift lakes. Sci Am. 280:64–69.
Streelman JT, Albertson RC, Kocher TD. 2003. Genome mapping of the orange blotch colour pattern in cichlid fishes. Mol Ecol. 12:2465–2471.
Sultmann H, Mayer WE, Figueroa F, Tichy H, Klein J. 1995. Phylogenetic analysis of cichlid fishes using nuclear DNA markers. Mol Biol Evol. 12:1033–1047.
Swanson WJ, Clark AG, Waldrip-Dail HM, Wolfner MF, Aquadro CF. 2001. Evolutionary EST analysis identifies rapidly evolving male reproductive proteins in Drosophila. Proc Natl Acad Sci USA. 98:7375–7379.
Swanson WJ, Vacquier VD. 2002. The rapid evolution of reproductive proteins. Nat Rev Genet. 3:137–144.
Tamura K, Nei M. 1993. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. Mol Biol Evol. 10:512–526.
Terai Y, Mayer WE, Klein J, Tichy H, Okada N. 2002. The effect of selection on a long wavelength-sensitive (lws) opsin gene of Lake Victoria cichlid fishes. Proc Natl Acad Sci USA. 99:15501–15506.
Terai Y, Morikawa N, Kawakami K, Okada N. 2002. Accelerated evolution of the surface amino acids in the WD-repeat domain encoded by the hagoromo gene in an explosively speciated lineage of East African cichlid fishes. Mol Biol Evol. 19:574–578.
Terai Y, Morikawa N, Okada N. 2002. The evolution of the prodomain of bone morphogenetic protein 4 (bmp4) in an explosively speciated lineage of East African cichlid fishes. Mol Biol Evol. 19:1628–1632.
Ting CT, Tsaur SC, Sun S, Browne WE, Chen YC, Patel NH, Wu CI. 2004. Gene duplication and speciation in Drosophila: evidence from the Odysseus locus. Proc Natl Acad Sci USA. 101:12232–12235.
Verheyen E, Salzburger W, Snoeks J, Meyer A. 2003. Origin of the superflock of cichlid fishes from Lake Victoria, East Africa. Science. 300:325–329.
Watanabe M, Kobayashi N, Shin-i T, Horiike T, Tateno Y, Kohara Y, Okada N. 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.
Wickler W. 1997. Sexually selected genital adornment and sperm packaging in species of Oreochromis (Teleostei: Cichlidae). Copeia. 1997:188–190.
Wong WS, Yang Z, Goldman N, Nielsen R. 2004. Accuracy and power of statistical methods for detecting adaptive evolution in protein coding sequences and for identifying positively selected sites. Genetics. 168:1041–1051.
Wu Q. 2005. Comparative genomics and diversifying selection of the clustered vertebrate protocadherin genes. Genetics. 169:2179–2188.
Yang Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. Comput Appl Biosci. 13: 555–556.
Yang Z, Bielawski JP. 2000. Statistical methods for detecting molecular adaptation. Trends Ecol Evol. 15: 496–503.
Yang Z, Nielsen R. 2000. Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. Mol Biol Evol. 17:32–43.
Yang Z, Nielsen R. 2002. Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. Mol Biol Evol. 19:908–917.
Yang Z, Wong WS, Nielsen R. 2005. Bayes empirical bayes inference of amino acid sites under positive selection. Mol Biol Evol. 22:1107–1118.

Billie Swalla, Associate Editor