International Institute for
Applied Systems Analysis
Schlossplatz 1
A-2361 Laxenburg, Austria

Tel: +43 2236 807 342
Fax: +43 2236 71313
E-mail: publications@iiasa.ac.at
Web: www.iiasa.ac.at

Interim Report **IR-11-026**

Repeated unidirectional introgression of nuclear and mitochondrial DNA between four congeneric Tanganyikan cichlids

Bruno Nevado (bruno.nevado@naturalsciences.be)
Varvara Fazalova (fazalova@iiasa.ac.at)
Thierry Backeljau (thierry.backeljau@ua.ac.be)
Mark Hanssens (mark.hanssens@africamuseum.be)
Erik Verheyen (erik.verheyen@ua.ac.be)

Approved by

Ulf Dieckmann
Program Leader, EEP

June 2011

Repeated unidirectional introgression of nuclear and mitochondrial DNA between four congeneric Tanganyikan cichlids

Manuscript submitted to *Molecular Biology and Evolution* as a research article

Running title: Interspecific geneflow in *Ophthalmotilapia* spp.

Bruno Nevado [1,2], Varvara Fazalova [3,4], Thierry Backeljau [1,2], Mark Hanssens [5] and Erik Verheyen [1,2]

1- Royal Belgian Institute of Natural Sciences, R. Vautier 29, 1000 Brussels, Belgium

2- Evolutionary Ecology Group, University of Antwerp, Groenenborgerlaan 171, B-2020 Antwerp, Belgium

3- Limnological Institute of the Siberian Branch of the Russian Academy of Sciences, Ulan-Batorskaya 3, 664033 Irkutsk, Russia

4- Evolution and Ecology Program, International Institute for Applied Systems Analysis, Schlossplatz 1, A-2361 Laxenburg, Austria

5-Royal Museum for Central Africa, Leuvensesteenweg 13, 3080 Tervuren, Belgium

Keywords: hybridization, introgression, mtDNA, nuclear DNA, *Ophthalmotilapia* spp.

This research was performed at the Royal Belgian Institute of Natural Sciences, Brussels, Belgium

Corresponding author: B. Nevado
Vertebrates Department
Royal Belgian Institute of Natural Sciences
Rue Vautier 29
1000 Brussels
BELGIUM
0032 2 627 44 28
bruno.nevado@naturalsciences.be

ABSTRACT:

With an increasing number of reported cases of hybridization and introgression, interspecific gene flow between animals has recently become a widely accepted and broadly studied phenomenon. In this study we examine patterns of hybridization and introgression in *Ophthalmotilapia* spp., a genus of cichlid fish from Lake Tanganyika, using mitochondrial and nuclear DNA from all four species in the genus and including specimens from over 800 kilometers of shoreline. These four species have very different, partially overlapping distribution ranges, thus allowing us to study in detail patterns of gene flow between sympatric and allopatric populations of the different species. We show that a significant proportion of individuals of the lake-wide distributed *O. nasuta* carry mitochondrial and/or nuclear DNA typical of other *Ophthalmotilapia* species. Strikingly, all such individuals were found in populations living in sympatry with each of the other *Ophthalmotilapia* species, strongly suggesting that this pattern originated by repeated and independent episodes of genetic exchange in different parts of the lake, with unidirectional introgression occurring into *O. nasuta*. Our analysis rejects the hypotheses that unidirectional introgression is caused by natural selection favoring heterospecific DNA, by skewed abundances of *Ophthalmotilapia* species, or by hybridization events occurring during a putative spatial expansion in *O. nasuta*. Instead, cytonuclear incompatibilities or asymmetric behavioral reproductive isolation seem to have driven repeated, unidirectional introgression of nuclear and mitochondrial DNA into *O. nasuta* in different parts of the lake.

INTRODUCTION:

With a growing number of hybridization events reported among animal species, the last decade has seen a shift in the way evolutionary biologists face hybridization amongst taxonomically valid species (e.g. Baack and Rieseberg 2007; Schwenk, Brede and Streit 2008). Recent research has shown that interspecific gene flow might be temporarily or spatially restricted, and may occur in some parts of the genome but not in others (reviewed in Baack and Rieseberg 2007), and the role of hybridization in animal speciation has been acknowledged (e.g. Mallet 2007). This shift in thinking became possible due to advances in theoretical research (Beerli and Felsenstein 1999; Beerli and Felsenstein 2001; Nielsen and Wakeley 2001; Hey and Nielsen 2004; Kuhner 2009), an increase in computing power, and the increasing availability of multilocus datasets for a variety of organisms

When studying interspecific hybridization, one often relies on the analysis of patterns of gene flow between two, usually sister, species (e.g. Carson and Dowling 2006; Di Candia and Routman 2007; Fitzpatrick et al. 2008; Nadachowska and Babik 2009). Strikingly, many such studies have found that following hybridization introgression of genetic material is typically unidirectional (e.g. McGuire et al. 2007; Alves et al. 2008; Plotner et al. 2008; Nevado et al. 2009; Keck and Near 2009). Different mechanisms have been suggested as responsible for this pattern, including natural selection (Ballard and Whitlock 2004; Nolte, Freyhof and Tautz 2006; Pfennig 2007; Alves et al. 2008; Plotner et al. 2008; Fitzpatrick et al. 2010), differences in the relative abundance of the involved species (Hubbs 1955; Wirtz 1999; Chan and Levin 2005; Carson and Dowling 2006; Linnen and Farrell 2007), cytonuclear incompatibilities or different survivability of reciprocal crosses (Bolnick et al. 2008),

or asymmetric behavioral reproductive isolation (Egger, Mattersdorfer and Sefc 2009).

Comparatively few studies have addressed patterns of hybridization between more than two species (Grant, Grant and Petren 2005; McDonald et al. 2008; Alves et al. 2008; Keck and Near 2009). These studies have increased our knowledge about interspecific hybridization by showing that two species may "hybridize indirectly" by common interaction with a third species (McDonald et al. 2008; Keck and Near 2009); and by showing that, in unidirectional introgression between multiple species, the donor species (Alves et al. 2008) or the recipient species (Keck and Near 2009) are often the same. These observations highlight the importance of an inclusive taxonomic sampling to properly understand patterns of gene flow amongst species, and suggest that intrinsic characteristics of the involved species may determine the outcome of hybridization events.

Particularly good predictors of the direction of introgression seem to be a species' distribution range and demographic history (Currat et al. 2008; Petit and Excoffier 2009). Under such scenario, individuals living at the distribution edge of a spatially expanding population become particularly susceptible to introgression of genetic material from species with which they come into contact during the range expansion. This could result in widespread species carrying genetic material of different species throughout its distribution range (e.g. Keck and Near 2009). The relative contribution of this phenomenon to the overall pattern of interspecific hybridization and introgression in animals depends on how many such cases are reported, and it thus seems timely to analyze groups of closely related species with different, overlapping distribution ranges.

Preliminary analyses revealed extensive sharing of mitochondrial DNA

(mtDNA) variation amongst the four species of the cichlid genus *Ophthalmotilapia*, part of the endemic Tanganyikan tribe Ectodini. If this sharing of genetic variation were due to interspecific hybridization, the *Ophthalmotilapia* genus would represent a particularly suitable group with which to address the role of a species' distribution range in the direction of introgression between potentially hybridizing species. This is because (i) the different species have different distribution ranges, which are partly overlapping (figure 1); (ii) the life history and ecological characteristics of the different species are very similar; and (iii) the Lake Tanganyika's species flock is old enough that it should allow discerning alternative scenarios of incomplete lineage sorting and introgression. The genus *Ophthalmotilapia* comprises four species: *O. boops*, *O. heterodonta*, *O. nasuta* and *O. ventralis* (Poll 1986); all these species feed on plankton and biocover and inhabit intermediate (sandy bottoms with rocks and boulders) shallow habitats (Poll 1986). All four species are sexually dimorphic, with males being more colorful and having extremely elongated pelvic fins with bright spatulated tassels at their tips (Hanssens, Snoeks and Verheyen 1999). The males also defend a breeding ground, to which females are attracted for spawning, and which varies from species to species. Following spawning, females mouthbrood the eggs and provide care for the fry (Nagoshi and Yanagisawa 1997). The four *Ophthalmotilapia* species show very different distribution patterns: *O. heterodonta* and *O. ventralis* have a complementary north-south distribution; *O. nasuta* is lake-wide distributed; and *O. boops* is restricted to a short stretch of shoreline in the south-eastern shore (Hanssens, Snoeks and Verheyen 1999). Given these dissimilar distribution ranges, we can analyze patterns of gene flow between both sympatric and allopatric populations of each species.

In this study, we address the role of interspecific gene flow in the evolutionary

history of a group of closely related species with very different and partly overlapping distribution ranges, by studying all the four species described in the genus *Ophthalmotilapia*. We use a mitochondrial gene (partial control region) and 9 nuclear microsatellites in order to answer the following questions: (i) can the genetic variation shared among species be attributed to hybridization (to the exclusion of alternative scenarios)?; (ii) if so, does the introgression of genetic material present a particular pattern, or is it random with respect to the species involved?; and (iii) if there is a particular pattern of introgression of genetic material, can any of the mechanisms typically invoked to explain such cases adequately account for the observed pattern?

MATERIAL AND METHODS:

Taxon and genetic sampling

Taxon sampling (supplementary table S1) included 44 *O. boops*, 36 *O. heterodonta*, 117 *O. nasuta* and 60 *O. ventralis* from 25 different localities across the lake (figure 1). Tissue samples collected in the field were kept in 80% ethanol until extraction of DNA following standard protocols (Quiagen DNeasy kit). The mitochondrial control region was amplified for all specimens using published protocols (Nevado et al. 2009) and the first most variable segment of the control region obtained by sequencing in the forward direction on a ABI 3130XL sequencer following the manufacturer's protocol.

For a subset of the specimens (178 individuals from all four species, see supplementary table S1) nine microsatellite loci were amplified: *UME002* and *UME003* (Parker and Kornfield 1996); *TmoM4*, *TmoM11*, *TmoM25* and *TmoM27* (Zardoya et al. 1996); *Pzeb1* and *Pzeb2* (van Oppen et al. 1997); and *OSU19D* (Wu,

Kung and Chow 1996) using the protocols detailed in supplementary table S2. The individuals in the microsatellite dataset were chosen so as to have representative population samples from several localities where patterns of gene flow between species were investigated in detail using mtDNA data (see below).

Phylogenetic analysis

Sequences were aligned with CLUSTALW v 1.83 (Thompson, Higgins and Gibson 1994) and the alignment checked by eye using the program SEAVIEW v 3.2 (Galtier, Gouy and Gautier 1996). Identical sequences were collapsed into haplotypes using the program collapse in TCS v 1.21 (Clement, Posada and Crandall 2000). Sites containing gaps were treated as a 5th character. JMODELTEST v 0.1.1 (Posada 2008) was used to select the best-fitting nucleotide substitution model (amongst three substitution type models) using both the Akaike Information Criterion (AIC, Akaike 1974) and the Bayesian Information Criterion (Schwarz 1978). Phylogenetic analysis was performed in MRBAYES v 3.1.2 (Huelsenbeck and Ronquist 2001; Ronquist and Huelsenbeck 2003) and in PHYML v 3.0 (Guindon and Gascuel 2003) using the previously selected nucleotide substitution model. For Bayesian inference two simultaneous runs (four chains each, temperature = 0.2) were sampled every 1000th generation for 10 million generations until the average split frequencies between the two runs reached a value smaller than 0.01. After removing a burn-in period and checking (by eye) its appropriateness, the remaining trees were summarized to obtain posterior probabilities for the nodes obtained. For maximum likelihood the parameters of the nucleotide substitution model and topology were optimized sequentially until no change in likelihood was found, and support for the resulting topology was obtained by performing 100 bootstrap replicates.

Analysis of gene flow using mtDNA

As the phylogenetic reconstructions showed that none of the four species is monophyletic (see results), and in order to distinguish between incomplete lineage sorting or interspecific gene flow as possible causes for the lack of monophyly, we used the Bayesian approach implemented in the program MIGRATE-N v 3.0.3 (Beerli and Felsenstein 1999; Beerli and Felsenstein 2001; Beerli 2006) to estimate amount and direction of gene flow between sympatric and allopatric populations of the different *Ophthalmotilapia* species. Our hypothesis was that if following separation the four *Ophthalmotilapia* species did not exchange genes, migration patterns between allopatric and sympatric populations (of different species) should be similar (e.g. Grant, Grant and Petren 2005). On the other hand, if some level of gene flow persisted following separation, migration patterns are likely to be different between sympatric and allopatric populations: there is no planktonic life-stage in *Ophthalmotilapia* species, and fertilization of eggs occurs inside the female's mouth, thus hybridization can only occur between sympatric populations, and gene flow should correspondingly be higher between such populations.

We treated individuals of each species collected in the same locality as representing a single population (insofar as at least 10 individuals were available). For *O. ventralis*, we treated individuals collected in localities 8 through 12 (figure 1) as belonging to the same population, due to the small number of individuals collected at each of these localities. Likewise, individuals of *O. heterodonta* from localities 19, 20 and 21 were treated as a single population. For each species, adequateness of the pooling of individuals was supported by non-significant F_{ST} values between these localities (data not shown). The populations defined in this step are hereafter named

according to the species and locality of origin (e.g. On_21 corresponds to *O. nasuta* individuals from locality 21; Oh_19-21 to *O. heterodonta* individuals from localities 19 to 21).

We performed all possible pairwise comparisons between populations from different species, in each analysis considering only two populations from different species. Parameters of the only nucleotide substitution model implemented in MIGRATE-N, F84 (Felsenstein and Churchill 1996) were estimated *a priori* in PAUP v 4.0.b10 (Sinauer Associates, Inc. Publishers) and kept constant throughout the runs. Priors for the analysis in MIGRATE-N were selected by performing preliminary runs, and selecting prior distributions that encompassed the complete posterior distributions for each parameter. Each pairwise population analysis was performed twice independently (100,000 steps as burn-in, followed by 500,000 states recorded every 100 steps). Results were checked by analyzing the shape of the posterior distributions for each parameter, the change in the parameters' values throughout each run (using TRACER v1.5, available from http://tree.bio.ed.ac.uk/software/tracer/) and by comparing the results from independent runs.

In case patterns of gene flow are different between sympatric and allopatric populations, and if this were due to interspecific genetic exchange, it would be interesting to gauge the time of such interspecific gene flow. With this purpose, we analyzed 100 datasets simulated under alternative scenarios of isolation and gene flow: the *Isolation* scenario (no gene flow after separation of the two populations, named scenario *I* hereafter), the *Migration* (constant amount of gene flow following separation, named *M*), the *Secondary Contact* (identical to the *Isolation* scenario, but with a recent period of gene flow, named *SC*) and the *Isolation with Gene Flow* scenarios (same as the *Isolation* scenario, but the two populations exchange genes at a

constant rate for a short period after the separation, denoted *IwGF*). We then qualitatively compared the patterns obtained under these simulated scenarios to the results obtained in the analysis of the real data.

The program SIMCOAL (Excoffier, Novembre and Schneider 2000) was used to simulate 100 datasets for each of four different scenarios of isolation and migration (*I, M, SC* and *IwGF*). The Kimura 2 parameter model of nucleotide substitution (Kimura 1980) was used for the simulations (amongst models implemented in SIMCOAL, this is the most similar to the best-fitting model selected for our real data) using a transition bias of 0.9 (resulting in transition/transversion ratios of approx. 13 as observed in our data). To mimic as closely as possible our real data, population sizes of the two populations were set to 150,000 individuals (inferred from thetas of 0.003 to 0.018 observed in our data) and kept constant (no strong growth/decline was detected in our real data). The substitution rate used for the simulations (2.8%) was the average of the values obtained using the method of Sturmbauer et al. (2001) in a dataset trimmed from the one used in Nevado et al. (2009) and corresponding to the first 500 bp of the control region (as used in this study). Time of separation of the two populations in each simulation was set to 500,000 years (from an average sequence divergence of 2-3% between the main mtDNA lineages in our data) and a 500 bp gene sequence was obtained for 20 individuals of each population at the end of the simulation. In the scenarios involving migration, a single individual was moved from population 1 to population 2 per generation (no migration in opposite direction): (i) for the entire 500,000 years (*M*), for the last 100,000 years of the simulation (*SC*), or for the first 100,000 years following separation (*IwGF*). For each scenario, simulated datasets were analyzed in migrate-n, and the results obtained (posterior distributions for migration values and for migration events through time) were averaged over the 100

datasets. The simulated datasets were also used in PAUP to build neighbor-joining trees, and the proportion of datasets resulting in reciprocal monophyletic relationships between the two simulated datasets was recorded for each simulation scenario.

Analysis of nested migration models in IMA

In order to obtain a more quantitative assessment of the support for alternative scenarios of gene flow between *Ophthalmotilapia* species, we performed a similar analysis of pairwise (sympatric and allopatric) congeneric populations, but used the program IMA (Hey and Nielsen 2007). For each pairwise comparison (involving *O. nasuta* populations), preliminary runs were performed to choose prior distributions for parameters of the model (population sizes of both current populations and of ancestral population, time of separation and migration amount between populations). We then performed two long runs for each analysis, with 100,000 burn-in steps followed by 100,000 sampled states (sampled every 100^{th} step). To fully explore the parameter space, we used a geometric heating scheme (30 chains, heating parameters 0.98 and 0.75). Convergence was checked by comparing results of the two independent runs. Resulting genealogies were then used to estimate posterior distributions for all parameters of interest, and to calculate the likelihood of the full isolation and migration model as well as that of nested sub-models (Hey and Nielsen 2007). We then performed likelihood ratio tests (LRT) to compare the full model (allowing migration in both directions) and three sub-models (no migration from population 1 to 2, no migration from population 2 to 1, and no migration in either direction). Significance of LRT was assessed by chi-square approximation, an approach which is statistically inappropriate because (i) the parameters of the nested sub-models were fixed at the boundary of the parameter space of the full model (Chernoff 1954) and

(ii) because the sites in the mtDNA gene do not represent independent observations (Nielsen and Wakeley 2001). We thus estimated, in addition to LRT, the AIC which calculates the relative support of the data for alternative models.

Neutrality tests and demographic reconstructions

As demographic histories of species and populations can affect patterns of hybridization and introgression, we estimated past demographic changes by estimating the following neutrality statistics in DNASP v 5.10 (Librado and Rozas 2009): Tajima's D (Tajima 1989), Fu and Li's D and F (Fu and Li 1993), Fu's Fs (Fu 1997) and Ramos-Onsins and Rozas R2 (Ramos-Onsins and Rozas 2002). Expectations for each test and significance of departure from these expectations were estimated with coalescent simulations (1000 replicates).

We further used the program BEAST v 1.5.3 (Drummond and Rambaut 2007) to reconstruct past demographic changes. We used the best fitting nucleotide substitution model (as selected by JMODELTEST) and with parameters of the model estimated by BEAST (except for the nucleotide frequencies, for which empirical values were used). We implemented a strict molecular clock, and priors for population size were obtained using the Bayesian skyride method (Minin, Bloomquist and Suchard 2008). All parameters' priors were set to their default values except for the alpha parameter of the gamma distribution (we used a uniform distribution between 0 and 100). Sampling was set to once every 1000^{th} steps for a minimum of 5 million steps and a maximum of 100 million steps (depending on datasets) in order to achieve Effective Sample Sizes (ESS) over 200. Though rarely, some parameters in a few runs exhibited ESS<200 (but higher than 100) after the maximum of 100 million steps, but sampling was not carried further than the 100 million steps. We checked for

convergence of independent runs using TRACER by plotting the change in likelihood values through each run, and by comparing results of two (or three when any ESS value was below 200) independent runs. Results from these runs were combined using the program LOGCOMBINER v 1.5.3 (part of the BEAST package) and demographic histories plotted.

Neutrality tests and demographic histories were estimated using the mitochondrial control region for each of the populations defined in the analysis of gene flow, and for each of the defined *O. nasuta* populations after removing suspected hybrid individuals (except for On_15, due to the low number of specimens carrying the nasuta I mtDNA lineage in this locality).

Microsatellites

Significance of deviations from Hardy-Weinberg equilibrium (HWE) in the nuclear dataset was estimated in GENEPOP v 4.0 (Raymond and Rousset 1995) by performing global (across all loci) heterozygosity excess and deficit tests for each locality and species. For each test, a Markov Chain (MC) was run for 10,000 steps (dememorization phase) followed by 20 batches of 5,000 iterations each. Linkage between loci was tested using the composite linkage disequilibrium test (Weir 1996) implemented in GENEPOP. Significance was assessed with MC sampling (same setting as above for HWE test).

The program STRUCTURE v 2.2 (Pritchard, Stephens and Donnelly 2000) was used to select the most likely number of clusters of individuals (K) in the dataset following the method of Evanno, Regnaut and Goudet (2005). For each value of K ($1 \leq K \leq 10$) three independent runs were sampled for 1 million generations following an initial burn-in period of 250,000 generations. Given the evidence in favor of

hybridization between *Ophthalmotilapia* species, we used the admixture model (individuals allowed to have mixed ancestries from different clusters) and assumed that allele frequencies are correlated among populations (i.e. that the allele frequencies in different clusters are likely to be similar due to migration or shared ancestry). Convergence of independent runs was checked graphically by plotting the likelihood values throughout each run, and by comparing likelihood values between the runs.

Analysis of gene flow using nuclear microsatellites

To investigate further patterns of gene flow between populations of different species, and to compare results from mtDNA with those from nuclear microsatellites, we performed a similar analysis of pairwise, sympatric and allopatric, populations of the different *Ophthalmotilapia* species, using the microsatellite dataset. As 3 of the 9 microsatellites used in this study have complex repetition motives (supplementary table S2), and given that a significant proportion of the alleles found in all microsatellites do not fit a stepwise mutation model (likely due to these microsatellites having been designed for species of a different cichlid tribe, the Tropheini), we could only use the infinite allele model (IA) as implemented in MIGRATE-N (IMA currently does not implement this model for microsatellites). Preliminary runs were used to choose adequate priors for theta and migration values, after which two independent runs were performed for each pairwise comparison (25,000 burn-in steps, followed by 100,000 states recorded every 100 steps, static heating scheme with 4 concurrent chains). Results of independent runs were checked for convergence by comparing resulting posterior distributions and estimated values of parameters.

RESULTS:

Phylogenetic analysis

We found 78 haplotypes amongst 257 control region sequences of *Ophthalmotilapia* spp., with eight of these haplotypes being shared between species (figure 2). Six of these eight haplotypes were shared between *O. nasuta* and one (or two) of the other *Ophthalmotilapia* species (figure 2), and these haplotypes were always found in sympatry or in nearby localities. We performed a permutation test to assess the probability that this pattern would occur by chance alone, in the following way: individuals of each of the species were randomly redistributed across localities, such that individuals carrying haplotypes shared amongst species could now be found in any locality. We then estimated, for each *O. nasuta* individual carrying a shared haplotype, the "geographic" distance between them and the individuals of the other species carrying the same haplotype. Note that the geographic distance was estimated by the number of localities between the two haplotypes, and as such does not accurately represent the real, physical distance between the individuals, but only the number of sampled localities between them. However, the same "geographic" distance was calculated for the empirical data, so this should not affect the result of this test. We performed 10,000 such permutations, and plot the resulting distributions of average and maximum distances obtained for these permutations in figure 3. The average and maximum number of distances estimated from the real data are represented by arrows in figure 3. It can be seen that both average and maximum number of distances in the real data do not fall in the ranges of distributions obtained in the permutations. This strongly suggests that the geographic distribution of shared

haplotypes observed in *Ophthalmotilapia* spp. is not random.

Both Akaike and Bayesian Information criteria implemented in JMODELTEST selected the Hasegawa-Kishino-Yano model of nucleotide substitution (Hasegawa, Kishino and Yano 1985) with a proportion of invariable sites and (gamma distributed) rate heterogeneity among the remaining sites (HKY+I+G). Bayesian Inference and Maximum Likelihood searches returned similar trees with the same (albeit often weakly supported) five main clades (figure 2, supplementary figure S3): one clade containing most (and exclusively) *O. nasuta* individuals, hereafter referred to as clade nasuta I; a second clade containing 17 *O. nasuta* individuals collected in localities 21 and 22 (clade nasuta II); a third clade containing most *O. boops* individuals (boops clade); a fourth clade containing most individuals of *O. ventralis* (ventralis clade); and a clade containing most individuals of *O. heterodonta* (heterodonta clade). The last two clades (ventralis and heterodonta) were not clearly separated, and two *O. heterodonta* individuals could not be assigned to either of these clades (figure 2). Interestingly, while several *O. nasuta* specimens clustered within the ventralis, the heterodonta or the boops clades, the two nasuta clades (nasuta I and nasuta II) contained exclusively *O. nasuta* individuals. Furthermore, *O. nasuta* individuals carrying mtDNA lineages typical of (i.e. clustering with) other species were always found within the distribution range of the latter species (inset figure 2).

Analysis of gene flow using mtDNA

We defined eight populations (four of *O. nasuta*, two of *O. ventralis* and one of each of the other species) and analyzed patterns of gene flow between all 21 possible interspecific population pairs. This included three sympatric population pairs (*O. nasuta* and each of the other three *Ophthalmotilapia* species) and 18 allopatric

comparisons (figure 4). In the analysis of *O. nasuta* and *O. boops* populations, it is evident that sympatric and allopatric populations of the two species exhibit different patterns of gene flow. In sympatric populations, the estimated amount of gene flow from *O. boops* into *O. nasuta* was much higher than between allopatric comparisons of populations of these species. On the other hand, the estimated amount of gene flow in opposite direction (from *O. nasuta* into *O. boops*) was similar across sympatric and allopatric comparisons (table 1, figure 4). A very similar pattern was found when analyzing *O. nasuta* and *O. ventralis* populations, with estimated amount of gene flow from *O. ventralis* into *O. nasuta* exhibiting much higher values in sympatry than in allopatry. Conversely, the estimated amount of gene flow between *O. nasuta* and *O. heterodonta* was generally low whether sympatric or allopatric populations were analyzed, with the highest amount of gene flow estimated from Oh_19-21 to On_15. Finally, amount of gene flow estimated between *O. boops* and *O. ventralis*, *O. boops* and *O. heterodonta* and *O. ventralis* and *O. heterodonta* (figure 4) was always low.

The results of our analysis of simulated datasets show that the *I* and *IwGF* scenarios can be distinguished from the *M* and *SC* scenarios by observing the shape of the posterior distribution for the amount of gene flow and the frequency histogram of migration events through time (figure 5). For the *I* and *IwGF* scenarios, posterior distributions for the amount of gene flow between populations (in both directions) peak at zero (i.e. the mode exhibited a very low value). For the *M* and *SC* scenarios, the posterior distribution for the amount of gene flow from population 1 to population 2 exhibit a mode at positive values (in the opposite direction, posterior distributions peak at zero). Similarly, the histogram of the frequency of migration events through time is different between the *M* and *SC*, and the *I* and *IwGF* scenarios. For the first two, the shape of the histogram resembles an exponential decay curve. For the latter

two the histogram of migration events through time resembles a sigmoid curve (in individual runs, the frequency of migration events exhibits positive modes, but given its variance across replicates the resulting summary distribution in figure 5 does not present a clear peak). Phylogenetic relationships were estimated in all 100 simulated datasets for each scenario. Most datasets obtained under the *I* (99) and under the *IwGF* (98) scenarios resulted in reciprocal monophyletic relationships between the two simulated populations. Conversely, in the *M* and *SC* scenarios almost all datasets resulted in polyphyletic relationships between simulated populations (98 in the *M* and 96 in the *SC* scenarios).

Analysis of nested migration models in IMA

In all runs performed in IMA, resulting posterior distributions for all parameters were very smooth and unimodal, with the exception of the parameter reflecting the time since separation of the two populations. Often, posterior distributions for this parameter exhibited long tails into high values, likely reflecting the long separation time of the two populations and as well the violation (in our data) of one of the major assumption of IMA: that the two populations under study are the closest related populations, and are not exchanging genes with other, unsampled populations. As our comparisons involved populations from different species, this assumption is certainly not met in our analysis, and thus the results obtained should be interpreted with caution. In this respect, it is worth noting that the results of independent runs were almost always very similar, not in the estimated likelihood of the models, but in the result of LRT and AIC analysis (table 2 and supplementary table S4).

For the comparisons involving *O. boops* and *O. nasuta,* LRTs were significant when comparing models with and without gene flow (into *O. nasuta*) between

sympatric populations (in locality 15). The AIC suggested that the best-fitting model for this sympatric comparison should include gene flow from *O. boops* into *O. nasuta* (and not in the opposite direction). For allopatric comparisons between *O. boops* and *O. nasuta* results mostly suggested absence of gene flow, with two exceptions: LRT was marginally significant in one of the runs comparing Ob_15 and On_18 (p=0.03); and AIC slightly favored a model including gene flow into *O. nasuta* in one of the runs comparing Ob_15 and On_18 (note that in both cases, these results were obtained only in 1 of the 2 independent runs of IMA, the other run suggesting no gene flow).

For the analysis involving *O. ventralis* and *O. nasuta*, the only comparison which returned consistent evidence for gene flow concerned sympatric populations Ov_8-12 and On_12: LRT tests in this case rejected models without gene flow into *O. nasuta*, and AIC selected a model including gene flow into (but not from) *O. nasuta*. For the comparisons involving the population On_15 and Ov_8-12 (or Ov_6), AIC suggested the best model to include gene flow into *O. nasuta*, however LRT tests were inconclusive. The remaining allopatric comparisons between these species suggest absence of gene flow (non-significant LRT and AIC selecting a model without gene flow).

In the analysis including *O. heterodonta* and *O. nasuta* populations, we recovered strong support for gene flow occurring between Oh_19-21 and On_15, with significant LRT results between models with and without gene flow into *O. nasuta*, and AIC favoring the model which includes gene flow into *O. nasuta*. None of the remaining comparisons returned significant LRT results, and AIC favored models without gene flow in two (out of three) comparisons.

Neutrality tests and demographic reconstructions

Most populations analyzed showed no signs of deviations from neutrality (table 3) with the exception of the populations Ob_15 (Tajima's D, Fu's Fs and R2 tests all significant). Likewise, the demographic histories of populations recovered using BEAST suggested rather stable population sizes for all populations analyzed, with the exception of Ob_15, which exhibited some growth (supplementary figure S5). Results obtained for *O. nasuta* were very similar whether all specimens, or only specimens carrying the nasuta I mtDNA clade, were analyzed (supplementary figure S5).

Microsatellites

Nine microsatellite loci were scored in 178 *Ophthalmotilapia* spp. (35 *O. boops*, 31 *O. heterodonta*, 76 *O. nasuta* and 36 *O. ventralis*) from 10 localities (supplementary table S1). Overall HWE was rejected for *O. nasuta* populations due to excess heterozygosity in locality 21 and due to deficit heterozygosity in localities 12 and 15 (table 4). None of the pairwise linkage disequilibrium tests was significant (supplementary table S6) suggesting that the nine loci are independent. The program STRUCTURE was used to estimate the likelihood of our nuclear dataset under a different number of clusters (K, between 1 and 10) using the admixture model with correlated allele frequencies, and this value was in turn used to estimate the statistic ΔK (Evanno, Regnaut and Goudet 2005). The likelihood rapidly increased between K=1 and K=3 or 4, and then plateaued for higher K values, while ΔK showed a clear peak at K=4 (figure 6). In the resulting bar plots with K=4 (figure 6) each cluster roughly corresponds to a species. The exceptions include the individuals of *O. ventralis* from the northern end of its distribution range, which cluster with *O. heterodonta*; and *O. nasuta* specimens from both ends of the analyzed distribution

range, with southernmost individuals always strongly clustering with *O. ventralis*, and northernmost individuals strongly clustering with either *O. ventralis* or *O. heterodonta* (in different runs).

Analysis of gene flow using nuclear microsatellites

Results of the pairwise population comparisons using the nuclear dataset are summarized in table 5 and supplementary figure S7. Posterior distributions for the amount of gene flow between populations in each analysis almost invariably returned unimodal, smooth distributions (only 2 posterior estimates showed bimodal distributions). However, independent runs for the same analysis often returned rather different results (table 5 and supplementary figure S7). Of the 64 estimated gene flow parameters, 12 exhibited posterior distributions with maximum frequency at 0, the remaining exhibiting a peak at positive values. The histogram of migration events through time (not shown) always resembled an exponential decay curve, with most of the posterior distribution situated at time=0. Contrarily to the mtDNA dataset, we did not recover any obvious relationship between estimates of gene flow and sympatric / allopatric comparisons (table 5).

DISCUSSION

Interspecific sharing of genetic variation: disentangling causes

The phylogenetic relationships of *Ophthalmotilapia* species show that none of the species is monophyletic. Even though species-specific mitochondrial lineages were retrieved, some haplotypes are shared between species, and several individuals carry mtDNA haplotypes typical of (i.e. genetically more similar to) other species' lineages.

Usual explanations for the lack of monophyly of closely related species include incorrect taxonomy, incomplete lineage sorting and interspecific hybridization (Funk and Omland 2003; McKay and Zink 2010).

A morphometric revision of the genus *Ophthalmotilapia*, with a full list of diagnostic characters, was given in Hanssens, Snoeks and Verheyen (1999). The most important and easily observable characters to distinguish the four species include the dentition, the width of the lower jaw and the number of scales between lateral lines (figure 7). *Ophthalmotilapia boops* can be distinguished from the three other *Ophthalmotilapia* species by its entirely tricuspid outer mandibular dentition (all other species have unicuspid outer oral teeth). *Ophthalmotilapia boops* (lower jaw width less than 28.0% of the Head Length, HL) and *O. nasuta* (less than 27.4 % HL) have a narrower lower jaw than *O. heterodonta* (more than 27.2 % HL) and *O. ventralis* (more than 24.1% HL). Due to allometry (the lower jaw width increases with increasing body size) there is a large overlap in the percentages for the entire size ranges of all four species. However, when these percentages are plotted against HL, almost all specimens can be assigned to either the narrow- or broad-mouthed species (figure 3 in Hanssens, Snoeks and Verheyen 1999). The anterior border of the lower jaw is also more rounded in *O. boops* and *O. nasuta*, while fairly straight in *O. heterodonta* and *O. ventralis* (the latter is illustrated in figure 7). The number of scales between the lateral lines can be used to distinguish between the narrow- and broad-mouthed species: *Ophthalmotilapia heterodonta* (two scales between the lateral lines) can be distinguished from *O. ventralis* (three scales); while *O. nasuta* (two scales) can also be distinguished from *O. boops* (three scales between the lateral lines). All individuals used in this study were carefully identified with vouchers kept for most specimens (supplementary table 1), and taxonomic misassignment seems unlikely to

account for the observed number of specimens carrying other species' DNA. For instance, 22 *O. nasuta* individuals cluster within the boops mtDNA clade, while distinguishing these two species is rather straightforward given the entirely tricuspid oral dentition of *O. boops*. Furthermore, there is no reason to expect a directional, biased taxonomic misidentification between species (i.e. that *O. heterodonta*, *O. boops* or *O. ventralis* would be more prone to be erroneously identified as *O. nasuta*, than vice-versa). Taken together, these observations suggest that the sharing of genetic material between *O. nasuta* and the remaining *Ophthalmotilapia* species cannot be adequately explained by incorrect taxonomic identification.

Incomplete lineage sorting and interspecific hybridization are often difficult to disentangle (Funk and Omland 2003; McKay and Zink 2010). In our study, several lines of evidence strongly supported the hybridization scenario over the incomplete lineage sorting. First, eight out of the 78 haplotypes found (c. 10%) were shared by at least two species (figure 2), which under the incomplete lineage sorting scenario would require extremely young species divergence times, which is unlikely given the amount of divergence observed both within and between the species. Second, the geographical pattern of haplotype sharing is clearly not random (figure 3), which would be expected under the incomplete lineage sorting scenario. Moreover, *O. nasuta* with mtDNA typical of other species (i.e. which in our phylogenetic reconstructions were solved within other species' mtDNA lineages) was only found in areas where *O. nasuta* occurs in sympatry with these other species (figure 2). Third, the number of haplotypes shared between species (or clustering within other species' mtDNA lineages) does not correlate with time since divergence. In fact the most closely related species *O. heterodonta* and *O. ventralis* (as revealed by both the mtDNA phylogeny and by the clustering of *O. ventralis* northern individuals with *O.*

heterodonta) exhibit relatively few such haplotypes when compared to the more distantly related *O. nasuta* and *O. ventralis* or *O. boops* (figure 2). And fourth, LRT and AIC analysis of nested migration models in IMA found consistent evidence (both LRT and AIC) for gene flow in sympatric (but not allopatric) comparisons between *O. boops* or *O. ventralis* and *O. nasuta* (table 2). For *O. heterodonta* and *O. nasuta*, the results are somewhat difficult to interpret, as both LRT and AIC results suggest significant gene flow between allopatric populations Oh_19-21 and On_15. This is due to a single *O. nasuta* individual found in locality 15 carrying *O. heterodonta* mtDNA (figure 2), and it is not clear why this would result in a significant inference of gene flow. It should be noted, however, that in all comparisons involving On_15, AIC selected a model with gene flow into *O. nasuta*, even when comparing this population to geographically very distant populations (table 2, supplementary table S4).

The analysis of simulated datasets also supports the claim that hybridization, and not incomplete lineage sorting, are responsible for the sharing of genetic material between *Ophthalmotilapia* species. In fact, the results obtained for the amount and timing of migration events (i.e. gene flow) when analyzing sympatric populations of *O. nasuta* and *O. boops* or *O. nasuta* and *O. ventralis* (figure 4 and table 1) is very similar to the results obtained for the simulated datasets under the *M* or *SC* scenarios (figure 5). Conversely, allopatric populations' comparisons almost invariably resulted in posterior distributions which resemble the ones obtained under the *I* and *IwGF* scenarios. Furthermore, out of 200 simulations under the *I* and *IwGF* scenarios, only three yielded non-monophyletic phylogenetic relationships between the two simulated populations (congruent with our own, real data). Conversely, for the two hybridization scenarios, the number of non-monophyletic results was 98/100 (*M*) and 96/100 (*SC*).

Therefore, our results provide strong support for the interspecific hybridization scenario. They suggest that reproductive isolation between the *Ophthalmotilapia* species is not complete, and in areas of sympatry individuals belonging to different species hybridize, leading to introgression of genetic material across species boundaries. A similar case has been reported in the Darwin's finches from the Galapagos Islands (Grant, Grant and Petren 2005), with microsatellite data showing that species living in sympatry on the same island often were more similar (genetically) than allopatric populations (from different islands) of the same species. In this work we use both nuclear as well as mtDNA, which allows us to make inferences regarding not only the presence/absence of introgression, but also the direction of these introgression events.

Patterns of introgression in Ophthalmotilapia spp.

We found that introgression of mtDNA in *Ophthalmotilapia* spp. occurs almost exclusively from *O. boops* or *O. ventralis* into sympatric *O. nasuta* individuals. The fact that *O. nasuta* individuals carry mtDNA haplotypes identical to other species' haplotypes, as well as some private haplotypes that cluster within other species mtDNA lineages suggests that hybridization has occurred in both historical and very recent times (possibly being ongoing at present). Furthermore, introgression seems to have taken place independently in different areas of the lake, and to have been rather extensive, with *O. nasuta* populations in some localities carrying a big proportion of mitochondrial DNA typical of other species (figure 2).

Contrarily to earlier studies on other animal groups (e.g. Alves et al. 2008; Keck and Near 2009), we recovered rather extensive introgression of nuclear DNA occurring concomitantly with introgression of mtDNA (figure 6). The results of our

analysis of nuclear microsatellites in MIGRATE-N was inconclusive, with some gene flow being recovered between almost all pairwise comparisons in both directions (table 5). To our knowledge, this is probably due to the complexity of the isolation with migration model, coupled with the relatively low information content of microsatellites using the IA model. This suggests that conclusions drawn using these 9 microsatellites should be taken with caution, particularly due to the high assignment probability of *O. nasuta* individuals to other *Ophthalmotilapia* species (figure 6). As such, the most-sound conclusions drawn in this study are based solely on mtDNA data, an approach with well-known caveats (e.g. Balloux 2010). In future work, the analysis herein performed could be complemented with genome-wide data including many nuclear genes, which could then be analyzed with better established nucleotide substitution models, and support the outcome of our analysis with mtDNA data. Regardless of these issues, our results of the clustering analysis performed in STRUCTURE suggest introgression of nuclear alleles from *O. ventralis* and *O. heterodonta* into *O. nasuta* (figure 6). This occurred in locality 12, with *O. ventralis*' nuclear DNA introgressing into *O. nasuta*; and in locality 22, where all *O. nasuta* individuals clustered with either *O. ventralis* or *O. heterodonta*. In the first case, introgression of nuclear DNA was accompanied by introgression of mtDNA (figure 2). In locality 22 (and 21), all 17 *O. nasuta* individuals collected carried one of four mtDNA haplotypes, which formed a monophyletic, strongly supported clade (clade nasuta II). Together with the ambiguous clustering in the nuclear DNA analysis, this suggests that these individuals represent an older hybridization event, between *O. nasuta* and *O. heterodonta*, or between *O. nasuta* and the ancestor of *O. ventralis* / *O. heterodonta*. If the latter hypothesis is correct, then the mtDNA lineage found in *O. nasuta* in these localities would represent a mtDNA fossil (*cf.* Bossu and Near 2009).

While introgression of some nuclear DNA is expected during introgression following hybridization, its traces are rarely recovered. This is typically attributed to the process of preferably unidirectional backcrosses of hybrids (and hybrids' progeny) with one of the parental species. Over a few generations, this should lead to the deletion of the other parental species' nuclear DNA traces. It should thus only be possible to recover nuclear DNA traces from both parental species in first (or early) generation hybrids. However, out of seven haplotypes clustering within the ventralis clade, and found in *O. nasuta* (figure 2), five were only found in the latter species, ruling out the possibility of these being early generation hybrids between these species. To our knowledge, the only explanation for the high proportion of introgressed nuclear DNA in *O. nasuta* several generations after the hybridization events (in the absence of selection favoring heterospecific nuclear DNA) is that *O. nasuta* females carrying heterospecific mtDNA successfully mate with *O. ventralis* or *O. heterodonta* males, resulting in a continuous introgression of nuclear DNA into *O. nasuta*.

Mechanisms responsible for introgression in Ophthalmotilapia spp.

Due to our inclusive taxonomic sampling of *Ophthalmotilapia* spp., our extensive geographic coverage, and the use of both nuclear and mtDNA markers, we are able to reject several of the potential mechanisms responsible for introgression in *Ophthalmotilapia* spp.

We uncovered extensive unidirectional introgression into the widespread *O. nasuta*, however this does not seem to have resulted from hybridization events occurring during a spatial expansion of this species. In fact, we did not find any traces of population expansion in any population of *O. nasuta* analyzed. Our results also

suggest that local populations of *O. nasuta* do not consistently outnumber sympatric populations of other species, and thus we can exclude the hypothesis of unidirectional introgression due to skewed abundances. Natural selection is also unable to account for the observed pattern of introgression, as this should result in a single mtDNA lineage introgressing into *O. nasuta* instead of several lineages in different areas. We are thus left with two possible explanations for the repeated unidirectional introgression of genetic material into *O. nasuta*: cyto-nuclear incompatibilities which affect reciprocal crosses differently, or asymmetric behavioral reproductive isolation. The two later hypotheses are hard to discern with our data, as they are expected to produce very similar patterns of interspecific sharing of genetic variation.

In favor of cyto-nuclear incompatibilities, it should be noted that *O. nasuta* females carrying *O. ventralis* mtDNA seem to have successfully mated with *O. ventralis* males, while no hybridization events seem to have involved *O. nasuta* females carrying nasuta I or nasuta II mtDNA lineages. If this is the case, our results provide further support for the claim by Keck and Near (2009) that the key factor influencing a species role as donor or recipient in unidirectional introgression events is the mtDNA lineage. On the other hand, asymmetric behavioral reproductive isolation through sexual selection could occur if females would favor heterospecific mates (e.g. Pfennig 2007), and could lead to introgression of nuclear genes responsible for sexual traits (e.g. Stein and Uy 2006). Hybrid viability studies and mate choice trials involving the different *Ophthalmotilapia* species would be needed to address this question. At any rate, our results show that intrinsic characteristics of species play an important role in deciding the fate of introgressing genes following hybridization, even in the absence of external factors shown to promote unidirectional introgression (such as the relative abundance of the species involved - e.g. Carson and

Dowling 2006; Linnen and Farrel 2007 - or their demographic and spatial dynamics - e.g. Currat et al. 2008; Petit and Excoffier 2009).

SUPPLEMENTARY MATERIAL

Table S1 - Taxon sampling.

Table S2 - Microsatellite protocols.

Figure S3 - Maximum-likelihood phylogenetic reconstruction.

Table S4 - LRT and AIC results with IMA.

Figure S5 - Reconstructed demographic histories of populations.

Table S6 - Results of linkage disequilibrium tests.

Figure S7 - Results of the microsatellite analysis with MIGRATE-N.

FUNDING

This work is part of B. Nevado PhD thesis, with funding from the Fundação para a Ciência e Tecnologia (grant SFRH/BD/17704/2004). Most fish used in this study were collected during different expeditions to Lake Tanganyika that were financed by the Belgian Science Policy (1992, 1995, 2001, 2006 and 2007, the two last in the context of MOLARCH - a EuroDIVERSITY Funded Collaborative Research Project) with support of the Leopold III Foundation for Nature Research and Nature Conservation.

ACKNOWLEDGMENTS

The authors would like to thank S. Koblmüller, K. Sefc, C. Sturmbauer and M. Virgilio for insightful discussions about the work herein presented. We are also thankful to Dr. J. Hey and two anonymous reviewers for useful comments on previous

versions of this manuscript, and Dr. P. Beerli for providing help with the program MIGRATE-N. P. Ngalande and the team at the Mpulungu Station of the Ministry of Agriculture and Cooperatives, Republic of Zambia and B. Ngatunga (Tanzanian Fisheries Research Institute) are acknowledged for their support to obtain the research permits and their assistance in the field. We are further grateful to L. Rüber for providing samples from R. D. Congo, and to Alain Reygel (RMCA) for providing the illustration for figure 7. Part of this work was carried out using the resources of the Computational Biology Service Unit from Cornell University, which is partially funded by Microsoft Corporation.

LITERATURE CITED

Akaike H. 1974. A new look at the statistical model identification. *IEEE Transactions on Automatic Control* 19: 716-723.

Alves PC, Melo-Ferreira J, Freitas H, Boursot P. 2008. The ubiquitous mountain hare mitochondria: multiple introgressive hybridization in hares, genus Lepus. *Philos Trans R Soc Lond B Biol Sci* 363: 2831-2839.

Baack EJ, Rieseberg LH. 2007. A genomic view of introgression and hybrid speciation. *Curr Opin Genet Dev* 17: 513-518.

Ballard JWO, Whitlock MC. 2004. The incomplete natural history of mitochondria. *Mol Ecol* 13: 729-744.

Balloux F. 2010. The worm in the fruit of the mitochondrial DNA tree. *Heredity* 104: 419-420.

Beerli P. 2006. Comparison of Bayesian and maximum-likelihood inference of population genetic parameters. *Bioinformatics* 22: 341-345.

Beerli P, Felsenstein J. 1999. Maximum-likelihood estimation of migration rates

and effective population numbers in two populations using a coalescent approach. *Genetics* 152: 763-773.

Beerli P, Felsenstein J. 2001. Maximum likelihood estimation of a migration matrix and effective population sizes in n subpopulations by using a coalescent approach. *Proc Natl Acad Sci U S A* 98: 4563-4568.

Bolnick DI, Turelli M, López-Fernández H, Wainwright PC, Near TJ. 2008. Accelerated mitochondrial evolution and "Darwin's corollary": asymmetric viability of reciprocal F1 hybrids in Centrarchid fishes. *Genetics* 178: 1037-1048.

Bossu CM, Near TJ. 2009. Gene Trees Reveal Repeated Instances of Mitochondrial DNA Introgression in Orangethroat Darters (Percidae: Etheostoma). *Syst Biol* 58: 114-129.

Carson EW, Dowling TE. 2006. Influence of hydrogeographic history and hybridization on the distribution of genetic variation in the pupfishes *Cyprinodon atrorus* and *C. bifasciatus*. *Mol Ecol* 15: 667-679.

Chan KM, Levin SA. 2005. Leaky prezygotic isolation and porous genomes: rapid introgression of maternally inherited DNA. *Evolution* 59: 720-729.

Chernoff H. 1954. On the distribution of the likelihood ratio. *Ann Math Statist* 25: 573-578.

Clement M, Posada D, Crandall KA. 2000. TCS: a computer program to estimate gene genealogies. *Mol Ecol* 9: 1657-1659.

Currat M, Ruedi M, Petit RJ, Excoffier L. 2008. The hidden side of invasions: massive introgression by local genes. *Evolution* 62: 1908-1920.

Di Candia MR, Routman EJ. 2007. Cytonuclear discordance across a leopard frog contact zone. *Mol Phylogenet Evol* 45: 564-575.

Drummond AJ, Rambaut A. 2007. BEAST: Bayesian evolutionary analysis by

sampling trees. *BMC Evol Biol* 7: 214.

Egger B, Mattersdorfer K, Sefc KM. 2009. Variable discrimination and asymmetric preferences in laboratory tests of reproductive isolation between cichlid colour morphs. *J Evol Biol* 23: 433-439.

Evanno G, Regnaut S, Goudet J. 2005. Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Mol Ecol* 14: 2611-2620.

Excoffier L, Novembre J, Schneider S. 2000. SIMCOAL: a general coalescent program for the simulation of molecular data in interconnected populations with arbitrary demography. *J Hered* 91: 506-509.

Felsenstein J, Churchill GA. 1996. A Hidden Markov Model approach to variation among sites in rate of evolution. *Mol Biol Evol* 13: 93-104.

Fitzpatrick BM, Johnson JR, Kump DK, Smith JJ, Voss SR, Shaffer HB. 2010. Rapid spread of invasive genes into a threatened native species. *Proc Natl Acad Sci USA 23:3606-10.*

Fitzpatrick BM, Placyk JS, Niemiller ML, Casper GS, Burghardt GM. 2008. Distinctiveness in the face of gene flow: hybridization between specialist and generalist gartersnakes. *Mol Ecol* 17: 4107-4117.

Fu YX. 1997. Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection. *Genetics* 147: 915-925.

Fu YX, Li WH. 1993. Statistical tests of neutrality of mutations. *Genetics* 133: 693-709.

Funk DJ, Omland KE. 2003. Species-level paraphyly and polyphyly: Frequency, Causes, and Consequences, with Insights from Animal Mitochondrial DNA. *Ann Rev Ecol Evol Syst* 34:397-423.

Galtier N, Gouy M, Gautier C. 1996. SEAVIEW and PHYLO_WIN: two graphic tools for sequence alignment and molecular phylogeny. *Comput Appl Biosci* 12: 543-548.

Grant PR, Grant BR, Petren K. 2005. Hybridization in the recent past. *Am Nat* 166: 56-67.

Guindon S, Gascuel O. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol* 52: 696-704.

Hanssens M, Snoeks J, Verheyen E. 1999. A morphometric revision of the genus Ophthalmotilapia (Teleostei, Cichlidae) from Lake Tanganyika (East Africa). *Zool J Linn Soc* 125:487-512.

Hasegawa M, Kishino H, Yano T. 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J Mol Evol* 22: 160-174.

Hey J, Nielsen R. 2004. Multilocus methods for estimating population sizes, migration rates and divergence time, with applications to the divergence of *Drosophila pseudoobscura* and *D. persimilis*. *Genetics* 167: 747-760.

Hey J, Nielsen R. 2007. Integration within the Felsenstein equation for improved Markov chain Monte Carlo methods in population genetics. *Proc Natl Acad Sci USA* 104: 2785-2790.

Hubbs CL. 1955. Hybridization between fish species in nature. *Syst Zool* 4: 1-20.

Huelsenbeck JP, Ronquist F. 2001. MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics* 17: 754-755.

Keck BP, Near TJ. 2009. Geographic and temporal aspects of mitochondrial replacement in Nothonotus darters (Teleostei: Percidae: Etheostomatinae). *Evolution* 64: 1410-1428.

Kimura M. 1980. A simple method for estimating evolutionary rates of base

substitutions through comparative studies of nucleotide sequences. *J Mol Evol* 16: 111-120.

Kuhner MK. 2009. Coalescent genealogy samplers: windows into population history. *Trends Ecol Evol* 24: 86-93.

Librado P, Rozas J. 2009. DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. *Bioinformatics* 25: 1451-1452.

Linnen CR, Farrell BD. 2007. Mitonuclear discordance is caused by rampant mitochondrial introgression in Neodiprion (Hymenoptera: Diprionidae) sawflies. *Evolution* 61: 1417-1438.

Mallet J. 2007. Hybrid speciation. *Nature* 446: 279-283.

McDonald DB, Parchman TL, Bower MR, Hubert WA, Rahel FJ. 2008. An introduced and a native vertebrate hybridize to form a genetic bridge to a second native species. *Proc Natl Acad Sci U S A* 105: 10837-10842.

McGuire JA, Linkem CW, Koo MS, Hutchison DW, Lappin AK, Orange DI, Lemos-Espinal J, Riddle BR, Jaeger JR. 2007. Mitochondrial introgression and incomplete lineage sorting through space and time: phylogenetics of crotaphytid lizards. *Evolution* 61: 2879-2897.

McKay BD, Zink RM. 2010. The causes of mitochondrial DNA gene tree paraphyly in birds. *Mol Phylogenet Evol* 54: 647-650.

Minin VN, Bloomquist EW, Suchard MA. 2008. Smooth skyride through a rough skyline: Bayesian coalescent-based inference of population dynamics. *Mol Biol Evol* 25: 1459-1471.

Nadachowska K, Babik W. 2009. Divergence in the face of gene flow: the case of two newts (Amphibia: Salamandridae). *Mol Biol Evol* 26: 829-841.

Nagoshi MT, Yanagisawa Y. 1997. Parental care patterns and growth and

survival of dependent offspring in cichlids. In: Kawanabe H, Hori M, Nagoshi M, editors. Fish Communities in Lake Tanganyika. Kyoto: Kyoto University Press. p. 175-192.

Nevado B, Koblmüller S, Sturmbauer C, Snoeks J, Usano-Alemany J, Verheyen E. 2009. Complete mitochondrial DNA replacement in a Lake Tanganyika cichlid fish. *Mol Ecol* 18: 4240-55.

Nielsen R, Wakeley J. 2001. Distinguishing migration from isolation: a Markov chain Monte Carlo approach. *Genetics* 158: 885-896.

Nolte AW, Freyhof J, Tautz D. 2006. When invaders meet locally adapted types: rapid moulding of hybrid zones between sculpins (Cottus, Pisces) in the Rhine system. *Mol Ecol* 15: 1983-1993.

Parker A, Kornfield I. 1996. Polygynandry in *Pseudotropheus zebra*, a cichlid fish from Lake Malawi. *Environ Biol Fish* 47: 345-352.

Petit RJ, Excoffier L. 2009. Gene flow and species delimitation. *Trends Ecol Evol* 24: 386-393.

Pfennig KS. 2007. Facultative mate choice drives adaptive hybridization. *Science* 318: 965-967.

Plotner J, Uzzell T, Beerli P, Spolsky C, Ohst T, Litvinchuk SN, Guex G-D, Reyer H-U, Hotz H. 2008. Widespread unidirectional transfer of mitochondrial DNA: a case in western Palaearctic water frogs. *J Evol Biol* 21: 668-681.

Poll, M. 1986. Classification des Cichlidae du lac Tanganika, Tribus, genres et espéces. Brussels: Académie Royale de Belgique.

Posada D. 2008. jModelTest: phylogenetic model averaging. *Mol Biol Evol* 25: 1253-1256.

Pritchard JK, Stephens M, Donnelly P. 2000. Inference of population structure

using multilocus genotype data. *Genetics* 155: 945-959.

Ramos-Onsins SE, Rozas J. 2002. Statistical properties of new neutrality tests against population growth. *Mol Biol Evol* 19: 2092-2100.

Raymond M, Rousset F. 1995. GENEPOP (version 1.2): population genetics software for exact tests and ecumenicism. *J Hered* 86: 248-249.

Ronquist F, Huelsenbeck JP. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* 19: 1572-1574.

Schwarz G. 1978. Estimating the dimension of a model. *The Annals of Statistics* 6: 461-464.

Schwenk K, Brede N, Streit B. 2008. Introduction. Extent, processes and evolutionary impact of interspecific hybridization in animals. *Philos Trans R Soc Lond B Biol Sci* 363: 2805-2811.

Stein AC, Uy JAC. 2006. Unidirectional introgression of a sexually selected trait across an avian hybrid zone: a role for female choice? *Evolution* 60: 1476-1485.

Sturmbauer C, Baric S, Salzburger W, Rüber L, Verheyen E. 2001. Lake level fluctuations synchronize genetic divergences of cichlid fishes in African lakes. *Mol Biol Evol* 18: 144-154.

Tajima F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* 123: 585-595.

Thompson JD, Higgins DG, Gibson TJ. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucl Acids Res* 22: 4673-4680.

van Oppen MJ, Rico C, Deutsch JC, Turner GF, Hewitt GM. 1997. Isolation and characterization of microsatellite loci in the cichlid fish *Pseudotropheus zebra*. *Mol*

Ecol 6: 387-388.

Weir BS. 1996. Genetic data analysis II: methods for discrete population genetic data. Sunderland, Mass: Sinauer Associates.

Wirtz P. 1999. Mother species-father species: unidirectional hybridization in animals with female choice. *Anim Behav* 58: 1-12.

Wu YM, Kung SS, Chow WY. 1996. Determination of relative abundance of splicing variants of Oreochromis glutamate receptors by quantitative reverse-transcriptase PCR. *FEBS Lett* 390: 157-160.

Zardoya R, Vollmer DM, Craddock C, Streelman JT, Karl S, Meyer A. 1996. Evolutionary conservation of microsatellite flanking regions and their use in resolving the phylogeny of cichlid fishes (Pisces: Perciformes). *Proc Biol Sci* 263: 1589-1598.

TABLES

Table 1: Average of the mode (from two runs) for the amount of gene flow estimated to (and from) *O. nasuta* populations with mtDNA.

	On_21	On_18	On_15	On_12
Ob_15	48.5 (3.8)	56.3 (3.8)	240.8 (4.7)	3.8 (3.8)
Ov_8-12	37.5 (3.8)	3.8 (3.8)	71.3 (3.8)	251.2 (3.8)
Ov_6	26.2 (3.8)	3.8 (3.8)	56.2 (3.8)	60.0 (3.8)
Oh_19-21	11.2 (3.8)	3.8 (26.2)	101.2 (3.8)	3.8 (3.8)

Filled cells highlight sympatric comparisons.

Table 2: Summary of the result of the analysis of nested migration models in IMA.

	On_21	On_18	On_15	On_12
Ob_15	m1=m2=0*	m1=m2=0**	**m1, m2=0**	m1=m2=0
Ov_8-12	m1=m2=0	m1=m2=0	m1, m2=0**	**m1, m2=0**
Ov_6	m1=m2=0	m1=m2=0	m1, m2=0**	m1=m2=0
Oh_19-21	m1=m2=0	m1=m2=0	**m1, m2=0**	m1=0, m2

m1 - gene flow into *O. nasuta*; m2 - gene flow from *O. nasuta.*
m1=0 means best model does not include gene flow into *O. nasuta*, m2=0 means it does not include gene flow from *O. nasuta* (as selected by AIC). Comparisons in bold denote significant LRT results between models with and without gene flow. Filled cells highlight sympatric comparisons.
*- runs returned different AIC-selected best models. **- runs returned different LRT results. For details, see supplementary table S4.

Table 3: Result of neutrality tests performed for each population of each species defined in the text.

Pop [A]	N [B]	Tajima's D	FL D [C]	FL F [D]	Fu's Fs [E]	R2 [F]
On_21	16	0.24 (0.69)	-0.5 (0.31)	-0.35 (0.33)	0.14 (0.54)	0.18 (0.58)
On_18	37	1.10 (0.90)	0.16 (0.54)	0.55 (0.73)	3.35 (0.90)	0.16 (0.88)
On_15	15	0.31 (0.65)	-0.18 (0.37)	-0.05 (0.43)	5.59 (0.98)	0.17 (0.77)
On_12	22	1.72 (0.98)	1.38 (0.98)	1.73 (0.99)	1.63 (0.79)	0.19 (0.99)
Oh_19-21	23	-0.91 (0.20)	0.13 (0.54)	-0.19 (0.35)	-1.77 (0.06)	0.09 (0.09)
Ob_15	26	-1.79 (0.01)*	-1.11 (0.14)	-1.53 (0.12)	-4.79 (0.00) *	0.08 (0.04) *
Ov_8-12	13	-0.21 (0.47)	-0.03 (0.47)	-0.09 (0.47)	-2.26 (0.09)	0.13 (0.28)
Ov_6	24	1.28 (0.92)	-0.26 (0.38)	0.22 (0.56)	2.75 (0.91)	0.18 (0.91)

[A] - Population name; [B] - number of sequences; [C] - Fu and Li's D; [D] - Fu and Li's F; [E] - Fu's Fs; [F] - Ramos-Onsins and Rozas R2. Significant deviations marked with *.

Table 4: Result (p-values) of Hardy-Weinberg Equilibrium tests (for heterozigosity deficit and excess) performed for the microsatellite dataset for each locality of each species.

Species	Locality	Excess	Deficit
O. boops	15	0.87	0.16
	18	0.16	0.82

O. heterodonta	15	0.08	0.93
	18	0.69	0.32
	19	0.77	0.24
	20	0.79	0.21
	21	0.87	0.13
O. nasuta	12	0.99	0.00 *
	15	0.99	0.01 *
	16	0.88	0.12
	18	0.67	0.39
	21	0.02*	0.98
O. ventralis	3	0.93	0.08
	6	0.89	0.18
	12	0.95	0.06
	15	0.96	0.12
	16	0.45	1.00

*- Denotes significant results

Table 5: Modes for the amount of gene flow estimated to (and from) *O. nasuta* populations using microsatellites.

	On_21	0n_18	0n_15	0n_12
Ob_15	23.5 (8.5) 19.5 (0.5)	0.5 (11.5) 29.5 (0.5)	4.5 (0.5) 0.5 (0.5)	114.5 (38.5) 42.5 (0.5)
Ov_12	0 (0) 76.5 (22.5)	32.5 (69.5) 8.5 (78.5)	20.5 (31.5) 22.5 (31.5)	33.5 (32.5) 49.5 (57.5)
Ov_6	32.5 (13.5) 16.5 (16.5)	0.5 (20.5) 0.5 (21.5)	5.5 (0.5) 0.5 (0.5)	91.5 (17.5) 26.5 (16.5)
Oh_19-21	24.5 (16.5) 36.5 (21.5)	5.5 (18.5) 19.5 (14.5)	22.5 (5.5) 4.5 (23.5)	30.5 (5.5) 31.5 (18.5)

Filled cells highlight sympatric comparisons. For each pairwise comparison, we show the results of two independent runs.

FIGURE LEGENDS:

Figure 1: Outline of Lake Tanganyika showing the 25 localities sampled in this study. The blue shades in the lake represent depth (darker areas are deeper). Inset shows approximate location of Lake Tanganyika in East Africa. Colored areas in the lake show distribution ranges analyzed in this study for *O. boops* (red), *O. heterodonta* (orange) and *O. ventralis* (blue). Note that our sampling did not include the complete distribution range of *O. heterodonta* (which inhabits most of the northern half of the

lake). Note also that *O. nasuta*'s distribution range encompasses the complete lake (not shown for clarity).

Figure 2: Phylogenetic relationships among 78 unique mtDNA haplotypes found in *Ophthalmotilapia* spp. Numbers below nodes denote posterior probabilities (PP), above nodes show bootstrap support (BS). Support not shown for branches with less than 0.5 (PP) or 50 (BS). Clade names follow from the text. Circles in front of branches of the tree represent number of individuals carrying each haplotype, and are colored according to the species where they were found (red for *O. boops*, orange for *O. heterodonta*, blue for *O. ventralis* and green for *O. nasuta*). Inset on the bottom shows the proportion of the different mtDNA lineages found in *O. nasuta* throughout the lake (number inside pie charts is the number of *O. nasuta* specimens collected in each locality; the proportion of the pie chart with each color represents the proportion of each mtDNA lineage found within *O. nasuta*'s individuals in each locality).

Figure 3: Result of the permutation analysis performed with haplotypes shared between species. The grey area represents the distribution of averages (NLOCaverage, upper panel) and maximum values (NLOCmaximum, lower panel) of the number of localities between haplotypes shared among species, obtained in 10,000 replicates where the geographic distribution of haplotypes within each species was random (x-axis is the number of localities between individuals carrying identical haplotypes, y-axis the number of observations). The arrows in both graphs signal the values of these two quantities (average and maximum) observed in the real data. As can be seen, these fall short of the complete distributions obtained by permutation, showing that the geographic distribution of shared haplotypes observed in our data is not random.

Figure 4: Results of the analysis performed in MIGRATE-N using different pairwise comparisons of populations from each species. Upper four rows represent comparisons between *O. nasuta* and each of the other *Ophthalmotilapia* species. Bottom two rows show pairwise comparisons between *O. boops*, *O. ventralis* and *O. heterodonta*. Inset "S" indicates comparisons between sympatric populations, remaining comparisons involve allopatric populations. Within each graph, gray lines represent posterior distributions for the amount of gene flow (bottom and left axes), black lines represent frequency histograms of migration events through time (upper and right axes). For both the estimated amount of gene flow and for the frequency histograms of migration events, dashed and solid lines depict the direction of gene flow (solid lines represent gene flow into the population depicted above the graphs, dashed lines gene flow into populations depicted on the left side). We show only results for one of two independent runs performed for each pairwise comparison, because the results were always identical in both runs. Lower and upper x axes are constant across graphs, and are only depicted on the lower and upper rows. For the amount of gene flow, a log scale is used on the y-axis to help visualization. The shaded area on the upper four rows represents the values for the amount of gene flow into *O. nasuta* which are inside the 95% highest-posterior credibility (HPC) set. Note how in the sympatric analysis involving *O. nasuta* and *O. ventralis* or *O. boops* the 95% HPC is shifted towards positive values and does not include zero.

Figure 5: Summary of the analysis in MIGRATE-N of simulated datasets obtained in SIMCOAL under different scenarios of isolation and gene flow (*Isolation*, *Isolation with Gene Flow*, *Migration* and *Secondary Contact*) between two populations that

diverged 500,000 years ago. Diagram on top of each panel depicts the timing and direction of gene flow events in each simulation (see material and methods for details). The graphs in each panel show the resulting distributions obtained for each parameter (in gray for the amount of gene flow, in black for the histogram of migration events through time) over 100 simulated datasets (black and dashed lines represent averages, dotted lines the 5 and 95% of the distribution) in each direction (shown in the upper-right of each graph). The layout of graphs follows from figure 3 (gene flow values depicted on left and bottom axes, frequency histogram of migration events on upper and right axes, scale of both x-axes kept constant across graphs, gene flow values depicted in log-scale).

Figure 6: Result of the analysis of nine microsatellite loci sampled in 178 *Ophthalmotilapia* spp. Upper graph shows the change in estimated likelihood (black line, represented on left y-axis) and in the statistic ΔK (gray, right y-axis) with increasing number of clusters assumed (K). The bar plot on the bottom was obtained with STRUCTURE when K=4 and using the admixture model with correlated allele frequencies. The results shown concern only one of the runs performed in STRUCTURE with K=4. Different runs of STRUCTURE with K=4 alternatively clustered *O. nasuta* individuals from locality 21 with *O. heterodonta* (assignment of all other individuals remained unchanged across runs). Each bar in the bar plot represents one individual, the color(s) in each bar representing the proportion of the individual's genome coming from each of the four assumed clusters. The bar plot is divided into four representing the taxonomic classification of the individuals into the four *Ophthalmotilapia* spp. Within each species thin black lines show different localities (denoted below plot).

Figure 7: Main taxonomic characters used in the identification of the different *Ophthalmotilapia* species. *Top panel*- Tricuspid (left) and unicuspid (right) teeth found in *Ophthalmotilapia* spp. *Middle panel*- Lower jaw width. This measurement is taken with the mouth open, to take this measurement the caliper is closed until it touches the edge of the lower jaw, which is at its widest anteriorly. *Bottom panel*- Transverse scale number. The transverse scales are counted at the origin or at the anterior part of the lower lateral line and only the non-perforated scales between the lateral lines are included.

Figure 1

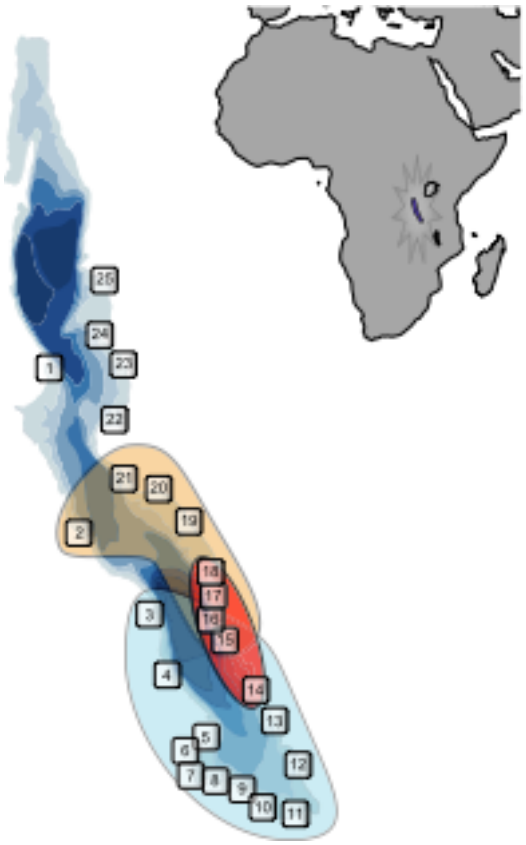

Figure 2

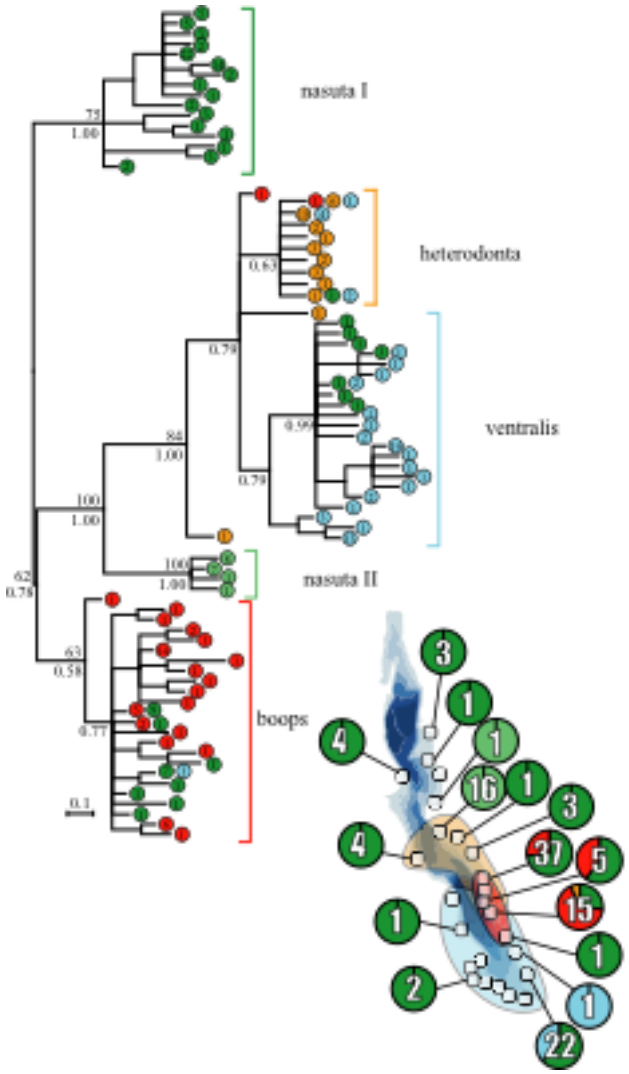
nasuta I
heterodonta
ventralis
nasuta II
boops
75
1.00
0.63
0.79
84
1.00
0.99
0.79
100
1.00
100
1.00
62
0.78
63
0.58
0.77
0.1
3
1
1
16
1
3
4
37
5
15
1
1
2
1
22

Figure 3

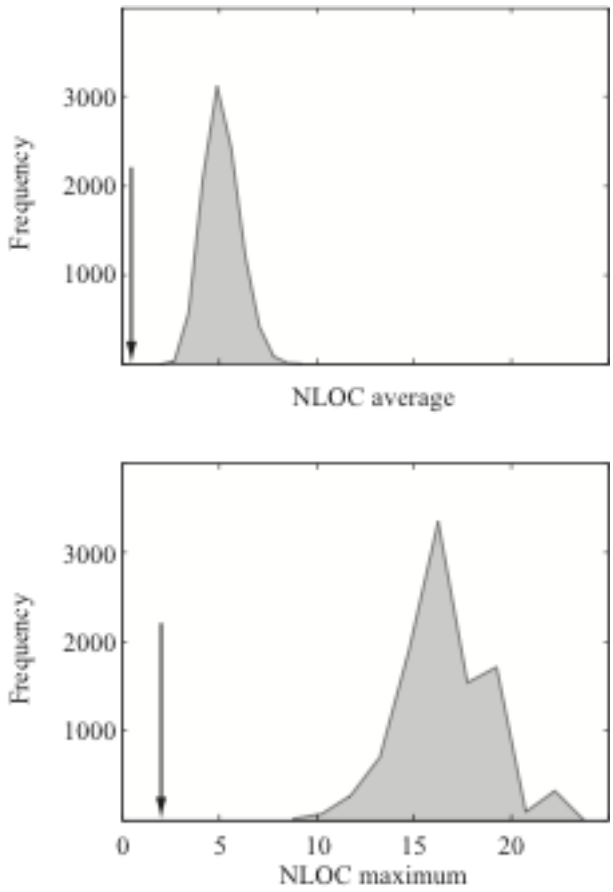
3000
2000
1000
Frequency
NLOC average
3000
2000
1000
Frequency
0
5
10
15
20
NLOC maximum

Figure 4

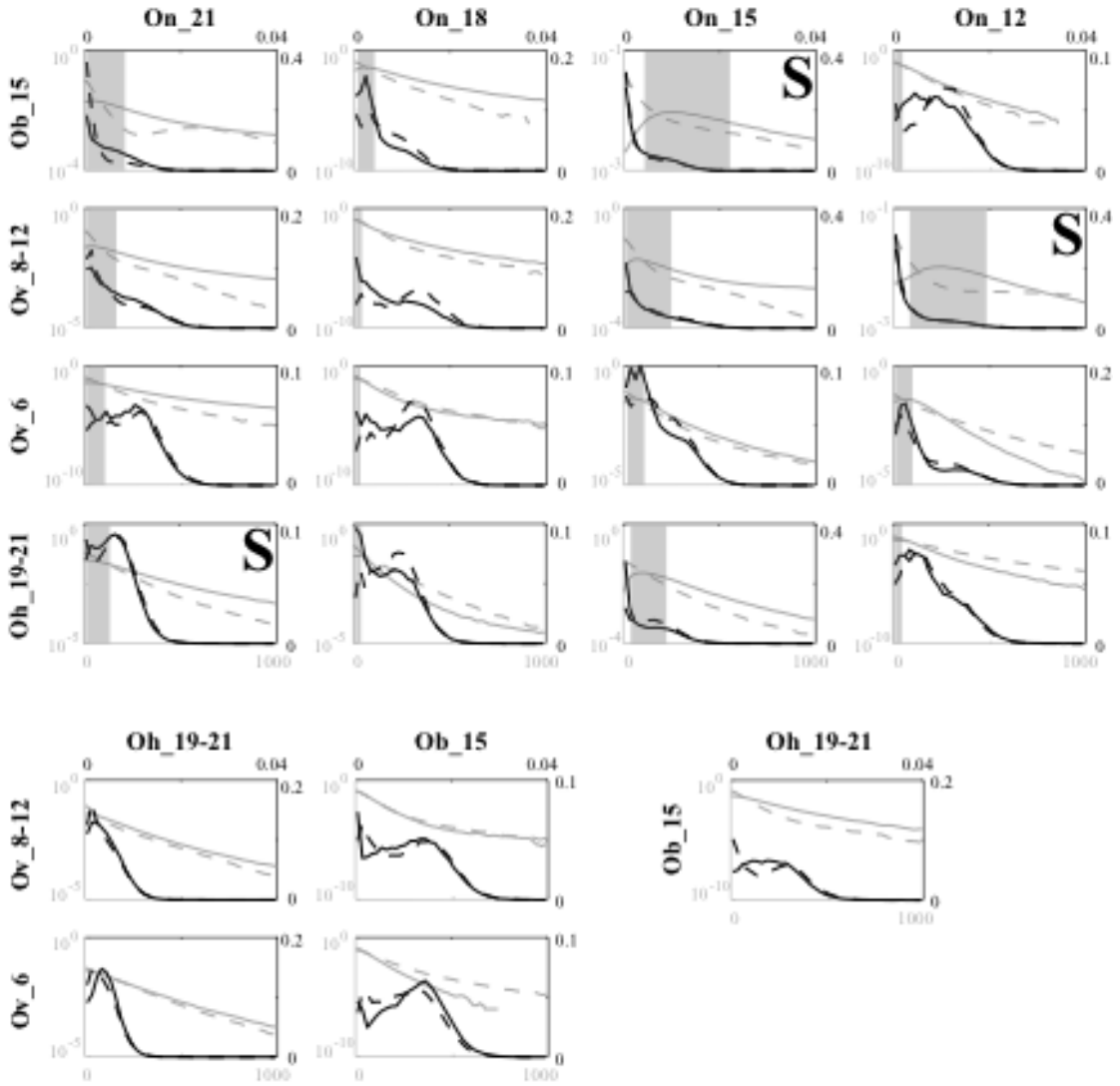
On_21
On_18
On_15
On_12
Ob_15
Ov_8-12
Ov_6
Oh_19-21
S
Oh_19-21
Ob_15
Ov_8-12
Ov_6
Oh_19-21
Ob_15

Figure 5

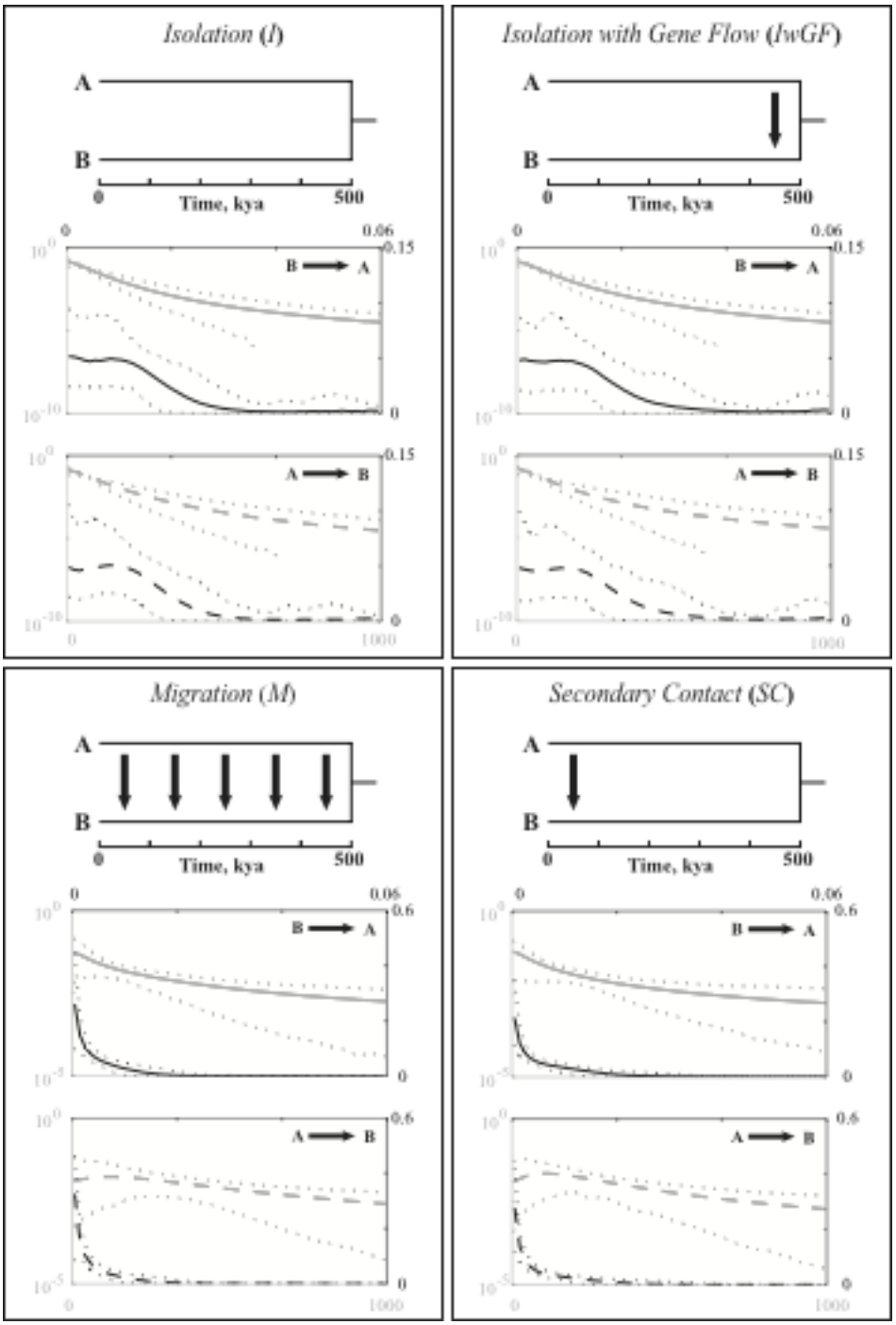
Isolation (I)
Isolation with Gene Flow (IwGF)
Migration (M)
Secondary Contact (SC)
A
B
0
Time, kya
500
B → A
A → B
0.06
0.15
0.6
1000

Figure 6

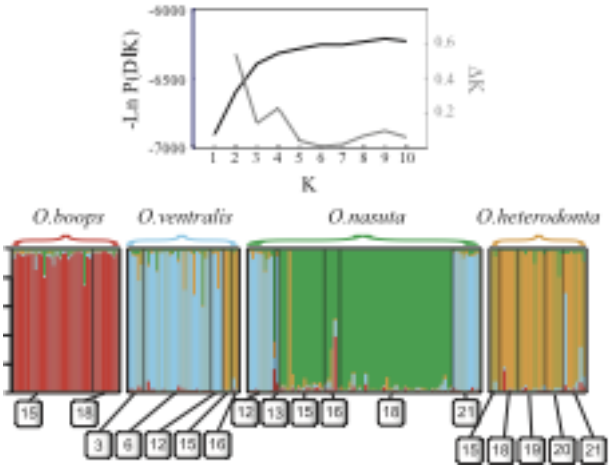

Figure 7