MOLECULAR ECOLOGY

Molecular Ecology (2016) **25**, 6143–6161 doi: 10.1111/mec.13767

Genomics of speciation and introgression in Princess cichlid fishes from Lake Tanganyika

HUGO F. GANTE,* MICHAEL MATSCHINER,† MARTIN MALMSTRØM,† KJETILL S. JAKOBSEN,† SISSEL JENTOFT†‡ and WALTER SALZBURGER*†

Zoological Institute, University of Basel, Vesalgasse 1, 4051 Basel, Switzerland, †Department of Biosciences, CEES (Centre for Ecological and Evolutionary Synthesis), University of Oslo, 0316 Oslo, Norway, ‡Department of Natural Sciences, University of Agder, 4604 Kristiansand, Norway

Abstract

How variation in the genome translates into biological diversity and new species originate has endured as the mystery of mysteries in evolutionary biology. African cichlid fishes are prime model systems to address speciation-related questions for their remarkable taxonomic and phenotypic diversity, and the possible role of gene flow in this process. Here, we capitalize on genome sequencing and phylogenomic analyses to address the relative impacts of incomplete lineage sorting, introgression and hybrid speciation in the *Neolamprologus savoryi*-complex (the 'Princess cichlids') from Lake Tanganyika. We present a time-calibrated species tree based on whole-genome sequences and provide strong evidence for incomplete lineage sorting in the early phases of diversification and multiple introgression events affecting different stages. Importantly, we find that the *Neolamprologus* chromosomes show centre-to-periphery biases in nucleotide diversity, sequence divergence, GC content, incomplete lineage sorting and rates of introgression, which are likely modulated by recombination density and linked selection. The detection of heterogeneous genomic landscapes has strong implications on the genomic mechanisms involved in speciation. Collinear chromosomal regions can be protected from gene flow and harbour incompatibility genes if they reside in lowly recombining regions, and coupling can evolve between non-physically linked genomic regions (chromosome centres in particular). Simultaneously, higher recombination towards chromosome peripheries makes these more dynamic, evolvable regions where adaptation polymorphisms have a fertile ground. Hence, differences in genome architecture could explain the levels of taxonomic and phenotypic diversity seen in taxa with collinear genomes and might have contributed to the spectacular cichlid diversity observed today.

Keywords: genomic landscapes, introgressive hybridization, linked selection, *Neolamprologus*, recombination bias, whole-genome sequencing

Received 7 January 2016; revision received 30 May 2016; accepted 11 July 2016

Introduction

The origin of new species – the process of speciation – has traditionally been viewed as a strictly bifurcating process, in which one biological species splits into two distinct entities as a consequence of the evolution of complete reproductive isolation between the newly emerged forms (*sensu* Mayr 1963; Coyne & Orr 2004). Once a topic of disagreement with botanists, the pervasiveness of introgressive hybridization, defined as the effective flow of genes beyond species boundaries (Anderson 1949; Barton & Bengtsson 1986), is no longer disputed by zoologists. The detection of hybridization and introgression between sister and nonsister species became commonplace with the introduction of molecular markers in evolutionary studies (Dowling & Secor 1997; Arnold 2006). In particular, advances in

Correspondence: Hugo F. Gante, Fax: +41 61 267 03 01; E-mail: hugo.gante@unibas.ch

genotyping methods and DNA sequencing have had a major impact on our understanding of the nature of species boundaries.

Earlier views of reproductive isolation as a property of species and, hence, of whole genomes gradually shifted towards a view where individual genes or specific regions of the genome are responsible for keeping species' distinctiveness (Wu 2001; Baack & Rieseberg 2007; Smadja & Butlin 2011; Feder *et al.* 2012). Furthermore, empirical studies are making it increasingly clear that interspecific gene flow is an integral part of the speciation process, with manifold potential evolutionary impacts (Abbott *et al.* 2013). For instance, while hybridization may lead to the fusion of species, it has been suggested that interspecific gene flow may in fact fuel rapid diversification by providing an abundant source of potentially adaptive genetic variation (Seehausen 2004; Berner & Salzburger 2015). In addition, new alleles introduced by introgression may not only be more abundant than alleles originating by mutation, but could also be more beneficial. This is likely because – in contrast to alleles arising *de novo* by mutation – introgressed alleles have already been exposed to natural selection, albeit in the genomic background of the introgression donor (Sætre 2013). Finally, introgression can also lead to the establishment of stable hybrid lineages that are reproductively isolated from parental species, following distinct evolutionary and ecological trajectories (Arnold 2006). Hybrid speciation through polyploidization has been shown to be particularly common in flowering plants and ferns, but a number of cases of homoploid hybrid species have also been documented, both in animals and in plants (DeMarais *et al.* 1992; Rieseberg *et al.* 2003; Baack & Rieseberg 2007; Mallet 2007; Mavárez & Linares 2008; Paun *et al.* 2009; Nolte & Tautz 2010; Salazar *et al.* 2010; Hermansen *et al.* 2014; Sousa-Santos *et al.* 2014; Trier *et al.* 2014).

Cline theory and hybrid zone studies have been extremely influential to our understanding of the dynamics of introgression. Coupling operating between isolation loci is maintained by linkage disequilibrium and contributes to limit gene flow when related species come into contact, whereas recombination acts in the opposite direction, eroding the association between physically linked loci (e.g. Haldane 1948; Barton 1983; Barton & Bengtsson 1986; Baird 1995; Gavrilets 2004; Bierne *et al.* 2011; Flaxman *et al.* 2014). Breaking down linkage with genes under negative selection allows introgression of neutral or positively selected loci that would otherwise be prevented from crossing species boundaries due to Hill–Robertson interference (Hill & Robertson 1966; Barton & Bengtsson 1986; Charlesworth 2009). Because selection acts particularly strongly on early-generation hybrids that carry maladaptive combinations of parental genes, a mechanistic role of hybrid 'filters' shapes the properties of introgressed chromosomal segments [or chromosomal blocks (Hanson 1959; Martinsen *et al.* 2001)]. Therefore, the key factor to *effective* interspecific gene flow is the ratio between selection and recombination in hybrids (Barton & Bengtsson 1986; Baird 1995).

Detection of introgression *vs.* incomplete lineage sorting and the characterization of hybridization remain major challenges (Abbott *et al.* 2013). Recent advances in genome-wide typing and whole-genome sequencing have provided unprecedented resolution to identify polymorphisms for adaptation to different niches or characterize the patterns of interspecific gene flow, including the detection of hybrid speciation, adaptive introgression and incomplete lineage sorting in a number of systems (Pollard *et al.* 2006; Kulathinal *et al.* 2009; Pardo-Diaz *et al.* 2012; Heliconius Genome Consortium 2012; Cui *et al.* 2013; Pease & Hahn 2013; Keller *et al.* 2013; Nadeau *et al.* 2013; Martin *et al.* 2013; Poelstra *et al.* 2014; Roesti *et al.* 2014, 2015; Gompert *et al.* 2014; Huerta-Sánchez *et al.* 2014; Berg *et al.* 2015; Eaton *et al.* 2015; Fontaine *et al.* 2015; Lamichhaney *et al.* 2015; Liu *et al.* 2015; Malinsky *et al.* 2015; Norris *et al.* 2015; but see Cruickshank & Hahn 2014). Rapidly diversifying groups are particularly prone to incomplete lineage sorting, and they are also the ones with the largest potential for consequential introgression and hybrid speciation, because of incipient isolation and exploitation of empty ecological niches. Therefore, the application of genomic data to study these groups will be of major importance for understanding the mechanisms of speciation and factors affecting interspecific gene flow.

Here, we employ whole-genome sequencing in combination with phylogenomic analyses to examine the interplay between genome architecture, speciation, and interspecific gene flow (isolation *vs.* introgression *vs.* incomplete lineage sorting) in a group of East African cichlid fishes. African cichlids constitute the largest extant vertebrate radiation, with several hundreds of species inhabiting lakes and associated river drainages (Turner *et al.* 2001; Salzburger 2009; Salzburger *et al.* 2014). Cichlids have thus become favourite model organisms for studying speciation mechanisms and factors that promote or retard diversification (Kocher 2004; Seehausen 2006; Gante & Salzburger 2012; Santos & Salzburger 2012; Wagner *et al.* 2012). In addition, five species of African cichlids from different lineages and geographical origins have had their genomes recently sequenced (Brawand *et al.* 2014). One of these lineages, the tribe Lamprologini, comprises almost half of the entire cichlid diversity in Lake Tanganyika (Gante & Salzburger 2012; Wagner *et al.* 2012). Lamprologines show complex ecological, morphological, trophic and behavioural attributes, representing a radiation within the cichlid

radiation (Stiassny 1997). Unlike other highly diverse cichlid lineages that display extreme levels of sexual dichromatism thought to promote their diversification, such as the radiations of haplochromine cichlids in lakes Malawi and Victoria (Wagner *et al.* 2012), most lamprologines are sexually monochromatic substrate spawners in which sexes share territorial defence and broodcare (Gante & Salzburger 2012). Furthermore, several examples of introgression have been reported in lamprologine cichlids (e.g. Schelly *et al.* 2006; Day *et al.* 2007; Koblmüller *et al.* 2007; Nevado *et al.* 2009; Sturmbauer *et al.* 2010), including a species of putative hybrid origin (Salzburger *et al.* 2002). Therefore, these taxa provide an extraordinary perspective on the reciprocal impacts of introgression and genome architecture on the evolution of an exceptionally diverse group of cichlids.

Materials and methods

Study system and specimen origin

Species of the *Neolamprologus savoryi*-complex, the 'Princess cichlids', are small (up to around 10 cm in standard length) fish endemic to Lake Tanganyika, eastern Africa. All species are cooperative breeders, in which the dominant, breeding couple is aided by up to 25 subordinate helpers in their tasks (territory maintenance, defence, broodcare), and the social group is organized in a strict linear hierarchy (Taborsky & Limberger 1981; Balshine *et al.* 2001). As a consequence, two species in this group, *N. brichardi* and *N. pulcher*, have emerged as important model systems for studying the evolution of cooperative breeding, so that substantial information on life history and behavioural traits has been documented (Wong & Balshine 2011). Social groups of Princess cichlids can be found in coastal rocky substrates of Lake Tanganyika between 3 and 50 m deep. The rocky substrate of Lake Tanganyika provides a territory with shelters and breeding grounds for adhesive eggs (Taborsky 1984; Heg *et al.* 2004). Each species is distributed around Lake Tanganyika and inhabits its rocky shores in a deme-like fashion, where allopatric populations of one species are often interspersed by populations of other species and by unfavourable (sandy) habitats (Konings 1998). In rare circumstances, up to five *Neolamprologus* species occur in sympatry (Büscher 1997).

Here, we investigated *Neolamprologus* species previously hypothesized to be involved in hybrid speciation (Salzburger *et al.* 2002) and for which nonmonophyletic mtDNA phylogenies have been found (Duftner *et al.* 2007). Namely, we made use of the published genome sequences of *N. brichardi* originating from the northeastern shore of Lake Tanganyika, in the Burundian–Kigoma area (Brawand *et al.* 2014), and sequenced whole genomes of inbred *N. gracilis* and *N. olivaceous* (hypothetical parental species), *N. marunguensis* (hypothetical hybrid species) all originating from the southwestern shore, in the Democratic Republic of Congo, and *N. pulcher* (polyphyletic with *N. brichardi*) originating from the southern shore, in Zambia (Table S1, Supporting information). Because samples originate from areas of allopatry, in several cases separated by hundreds of kilometres of shoreline, they represent the evolutionary outcome of ancient gene flow (hybridization and introgression) events, rather than present-day hybrid zone dynamics.

Sample processing and whole-genome sequencing

High-quality genomic DNA was isolated from fresh fin tissue of one female individual per species using Qiagen DNeasy® Blood and Tissue Kit, diluted to 11–14 ng/μL (100 μL) with Qiagen Elution Buffer (Qiagen) and fragmented to ~300 bp by sonication on a Bioruptor® Next Gen. All libraries were constructed following Illumina TruSeq Sample Prep v2 Low-Throughput Protocol, including agarose gel size selection, and sequenced on Illumina HiSeq 2500. This yielded 157.1–180.6 million PE reads of length 101 bases each, and an average insert size of 269–311 bases. To assess the sequence quality and gene space completeness of the newly sequenced *Neolamprologus* genomes, we first assembled each genome *de novo* using the CELERA ASSEMBLER package (Miller *et al.* 2008). We then used the software CEGMA version 2.4.010312 (Parra *et al.* 2007, 2009) to investigate the presence ('partial' or 'complete') of 248 conserved eukaryotic genes with few paralogs. Assembly coverage (18.5–21.4×) and the numbers of 'partial' and 'complete' genes recovered for each species are given in Table S2 (Supporting information).

Whole-genome alignments

Taking advantage of high collinearity among cichlid genomes (Brawand *et al.* 2014), we used the Illumina raw reads to align sequences of all five *Neolamprologus* species against the high-quality *Oreochromis niloticus* linkage groups (i.e. chromosomes). The Lake Malawi cichlid *Metriaclima zebra* was included to break the longer branch to tilapia and improve phylogenetic reconstruction accuracy. Comparable sequence data sets for *N. brichardi* and *M. zebra* (Brawand *et al.* 2014) were chosen based on similarity in sequencing technology, read length, insert size, and the number of reads, and fastq files for these data sets were downloaded from the European Bioinformatics Institute (www.ebi.ac.uk/ena/data/view/SRR077345 and www.ebi.ac.uk/ena/data/view/SRR077295, respectively). Sequences for the 23

O. niloticus (Orenil1.1) linkage groups were downloaded from NCBI (http://www.ncbi.nlm.nih.gov/assembly/GCF_000188235.2/) and merged into a single file, excluding any unplaced scaffolds and contigs. Given the good and similar quality of the new genomes, the raw, unfiltered reads for each species were then mapped to the indexed *O. niloticus* reference genome with BWA version 0.7.9a (Li & Durbin 2009) and a sorted BAM file was produced with SAMTOOLS version 1.0 (Li *et al.* 2009). The MARKDUPLICATES software from the PICARD-TOOLS version 1.107 package (http://broadinstitute.github.io/picard/) was used to identify and filter out duplicate reads from PCR. The deduplicated BAM files were then indexed, and a pileup file was created for each species. Per-base alignment qualities were recalculated and reads with mapping quality below 30 and base phred score below 30 were excluded while allowing 'orphan reads' (no read pair) to be kept in the file. The mean mapping coverage was 14.2× for *N. brichardi*, 16.30× for *M. zebra* and 15.0×–17.2× for the four newly sequenced *Neolamprologus* species (Figure S1, Supporting information). Importantly, only between 2.9% and 6.6% of the sites had a low but non-zero coverage between 1× and 4×, suggesting that few heterozygous sites were misidentified as homozygous. The pileup files were used to identify the variants and generate one VCF file per species with BCFTOOLS version 1.1 (invoking the 'call-c' option) (Li 2011). Special consideration was taken to circumvent alignment issues relating to indels. Species-specific insertions in the mapped genomes would not be phylogenetically informative for this type of analysis and would also be disregarded by the nature of the mapping approach. To maintain a consistent alignment length across all species, short deletions in the mapped species compared to *O. niloticus* were replaced with missing data ('n') in the VCF files. The VCF files were then converted into fastq file with VCFTOOLS version 0.1.12b (Danecek *et al.* 2011) and subsequently to fasta format using SEQTK v.1.0-r75 (https://github.com/lh3/seqtk), whereby heterozygous sites were incorporated by using IUPAC ambiguity codes. We generated alignment files for each linkage group by combining the fasta entries of *O. niloticus*, *M. zebra* and the five *Neolamprologus* species. All alignments were then manually inspected and alignment ends were trimmed (on the order of a few hundred bases due to repeats) to remove terminal regions in which none of the six mapped species had coverage.

Identification of chromosomal segments supporting different local topologies

Phylogenetic relationships of genomic regions may differ from the species tree due to incomplete lineage sorting and introgression. To test whether this is the case for the five investigated genomes of *Neolamprologus*, and to detect breakpoints between genomic segments supporting different local topologies, we used the machine-learning approach implemented in SAGUARO version 0.1 (Zamani *et al.* 2013). SAGUARO does not require any a priori phylogenetic hypotheses, but infers similarity matrices and segment boundaries strictly from the genomic data. The analysis was performed jointly for alignments of all linkage groups, to improve inference of similarity matrices, which can map to multiple regions on different linkage groups. We constrained SAGUARO to use only nucleotide positions with less than 50% missing data and a minor allele frequency of two; otherwise, default parameters were applied.

We produced local alignments for all segments identified by SAGUARO, and inferred maximum-likelihood (ML) phylogenies for each alignment using RAXML version 8.1.17 (Stamatakis 2014). Due to the small number of taxa, we applied a GTR model (Tavaré 1986) of sequence evolution with no rate heterogeneity among sites (RAXML option-V) in each analysis. Outgroup positions were assumed for *O. niloticus* and *M. zebra*, relative to the five *Neolamprologus* species. Following ML analyses, topological support was assessed using bootstrapping (Felsenstein 1985) with the autoMRE bootstopping criterion. To identify a subset of segments with the most reliably inferred topologies, we filtered based on read coverage, alignment properties and tree support. Specifically, we included only segments with a minimum length of 500 000 bp, a maximum proportion of 50% of missing data per species, a mean phylogenetic bootstrap support of 80 (averaged over nodes) and a maximum proportion of 50% of alignment positions with low read coverage (less than 3× in one or more of the newly sequenced genomes). By filtering out low-coverage regions, we likely also reduced the number of heterozygous sites that falsely appeared as homozygous due to null alleles. A total of 224 segments distributed across all linkage groups fulfilled these criteria, to which we will refer as 'most reliable segments'. These contained a minimum of 4311 parsimony-informative sites per segment (median: 8258.5; mean: 11 477.4).

Inference of time-calibrated local phylogenies

To infer the root position and sequence of divergence events under a molecular clock model, we performed Bayesian phylogenetic analyses with the software BEAST version 2.2.0 (Bouckaert *et al.* 2014) for alignment blocks of 500 000 bp sampled from all 'most reliable segments', totalling 426 blocks (32.4% of all alignment sites). Each alignment block was analysed with three different models: (i) a strict molecular clock model

without rate variation across sites (GTR; 50 million Markov chain Monte Carlo (MCMC) generations); (ii) a strict molecular clock model with gamma-distributed rate variation across sites (GTR+Gamma; 100 million MCMC generations); and (iii) a relaxed-clock model with uncorrelated log-normal branch rate variation (100 million MCMC generations using the UCLN model of Drummond *et al.* 2006). We assumed outgroup positions for *O. niloticus* and *M. zebra* and used age estimates for the divergences of these two taxa according to McMahan *et al.* (2013) to time-calibrate the phylogeny. Despite a comparatively low number of temporal constraints used by McMahan *et al.* (2013), their estimated timeline of cichlid diversification is intermediate between those of other studies (Azuma *et al.* 2008; Friedman *et al.* 2013) and consistent with an endemic Lake Tanganyika radiation, which is supported by geomorphological evidence and by broader phylogenetic hypotheses of Lake Tanganyika cichlids (Cohen *et al.* 1993; Salzburger *et al.* 2005, 2014; Meyer *et al.* 2016). The divergence between Oreochromini (including *O. niloticus*) and Austrotilapiini (including *M. zebra* and *Neolamprologus*) was estimated by McMahan *et al.* (2013) to have occurred at 17.3–31.6 Ma, which we here modelled with a log-normal divergence prior distribution with offset 8.2 Ma, mean in real space 15.067 Ma and a standard deviation in log space of 0.24. In addition, the divergence event separating the Haplochromini (including *M. zebra*) and the Lamprologini (including *Neolamprologus*) was estimated at 12.7–24.7 Ma by McMahan *et al.* (2013), which we approximated with a log-normal prior distribution with offset 5.1 Ma, mean in real space 12.6 Ma and a standard deviation in log space of 0.24. Stationarity of MCMC chains was assessed by inspecting parameter traces and their effective sample sizes with TRACER version 1.6 (Rambaut *et al.* 2014). To test whether stationarity also indicated convergence, we repeated all analyses with different starting positions of the MCMC chain. The relative fit of the three models of molecular evolution was assessed *a posteriori* based on an Akaike information criterion through Markov chain Monte Carlo (AICM; Baele *et al.* 2012) analysis as implemented in TRACER version 1.6. After discarding the first 20% of posterior tree samples as burn-in, the remaining samples were used to calculate segment-specific maximum clade credibility (MCC) trees, with node heights set to mean posterior age estimates.

Three-taxon trees and genome-wide Patterson's D statistics

To disentangle the effects of incomplete lineage sorting and introgression, which could both lead to incongruence between local phylogenies, we used local and genome-wide approaches. For the former, we calculated the numbers of phylogenies supporting distinct rooted topologies among sets of three *Neolamprologus* species. If the true species tree of three taxa A, B and C was ((A,B),C) and local phylogenies were affected by incomplete lineage sorting, we would expect to observe the alternative topologies ((A,C),B) and ((B,C),A) in roughly similar proportions. However, if local phylogenies differ due to introgression between C and either A or B, we would expect to see only one of the alternative topologies (unless introgression occurred in both A and B). Thus, frequencies of the observed topologies among three taxa can inform about the occurrence of incomplete lineage sorting and introgression.

In a genome-wide approach, we tested for excess of shared allelic variants in four-taxon comparisons using Patterson's *D* statistic (Green *et al.* 2010). This is the appropriate scale to use the *D* statistic, as the potential for erratic behaviour has been reported when applied to small genomic regions with low effective population size (Martin *et al.* 2015b). Assuming a species tree (((*P1*, *P2*),*P3*),*O*) of three species of *Neolamprologus* (*P1*–*P3*) and *Metriaclima zebra* as the outgroup (*O*), we counted the number of biallelic sites at which *P1* and *P3* ('BABA' sites) or *P2* and *P3* ('ABBA' sites) share the derived state. Under a null hypothesis of strict bifurcation without introgression, the frequencies of ABBA and BABA sites across the genome should not be significantly different. Thus, the reasoning behind this test is similar to the comparison of frequencies of segment-specific topologies described above. In contrast to the segment-specific topology frequency comparison, however, this test does not depend on correct identification of segment boundaries and can be applied to the full genome alignment. Moreover, significance of the excess of ABBA or BABA sites can be assessed by calculation of the standard error and *z*-scores through a standard block jackknife procedure (Durand *et al.* 2011). Based on the results of our local phylogeny inference, we calculated the *D* statistic for ten sets of three *Neolamprologus* taxa (plus outgroup *Metriaclima zebra*) so that *N. marunguensis* was considered sister to two other species of *Neolamprologus*, whenever it was included. For comparisons excluding *N. marunguensis*, *N. gracilis* was assumed at the same position, and for one comparison excluding both *N. marunguensis* and *N. gracilis*, *N. brichardi* was used as the sister to *N. olivaceous* and *N. pulcher*. The length of jackknife blocks was set to 100 000 bp to account for nonindependence of sites due to linkage disequilibrium. Because we are interesting in ancient introgression events, only fixed SNPs were used in the calculations. These constitute the vast majority of the available data (Table S3, Supporting information).

Maximum-likelihood tests of introgression

The observed frequencies of three-taxon gene trees and genome-wide *D* statistics allowed us to formulate explicit hypotheses of introgression and incomplete lineage sorting, which we then tested in a maximum-likelihood framework. These included three introgression events: between *N. brichardi* and *N. pulcher*, between *N. marunguensis* and the common ancestor of *N. pulcher* and *N. olivaceous* and between *N. marunguensis* and *N. gracilis*. We used the CalGTProb method (Yu *et al.* 2014) implemented in PhyloNet v.3.5.5 (Than *et al.* 2008) to first assess the likelihood of an introgression-free species tree, given the set of 426 time-calibrated MCC topologies produced with BEAST. We then compared this likelihood to likelihoods of species trees with one to three reticulation edges, representing introgression events. Because the CalGTProb method assumes reticulation edges to be directional, indicating the donor and recipient of introgression, we tested both directions for each added reticulation edge and retained the edge direction that resulted in a better likelihood. As part of the analysis, weights of reticulation edges are estimated, which correspond to the proportion of the recipient's genome resulting from introgression (Yu *et al.* 2014).

Timing of species divergences and introgression events

We further used the set of 426 time-calibrated local MCC phylogenies to infer divergence times and ages of introgression events. To do so, we sorted phylogenies into four partially overlapping subsets, for (i) phylogenies resulting from introgression from *N. brichardi* to *N. pulcher*, (ii) phylogenies affected by introgression from *N. marunguensis* to the common ancestor of *N. olivaceous* and *N. pulcher*, (iii) phylogenies likely to reflect introgression from *N. marunguensis* to *N. gracilis* and (iv) presumably introgression-free phylogenies. Specifically, we included all phylogenies with rooted topology IDs R6, R7, R10, R11, R13, R17 and R18 (see Table S4, Supporting information), in which *N. brichardi* appeared as the sister to *N. pulcher*, in subset (i), and phylogenies with topology IDs R4, R8, R9, R10, R12, R15, R16, R17, R18 and R19, in which *N. marunguensis* was the closest relative to *N. olivaceous* except *N. pulcher*, in subset (ii). Local phylogenies of the three taxa *N. marunguensis*, *N. gracilis* and *N. brichardi* appear to have been influenced by both incomplete lineage sorting and introgression. In the three-taxon comparison, *N. marunguensis* and *N. gracilis* were found as sister species in 149 local phylogenies, whereas *N. marunguensis* formed a clade with *N. brichardi* in 92 phylogenies. It is likely that the former 149 phylogenies reflect a mix of incomplete lineage sorting and introgression, while the latter 92 phylogenies result from incomplete lineage sorting alone (Table S5 and Figure S2, Supporting information). As incomplete lineage sorting should affect both alternative topologies equally, we assume that the sister group relation between *N. marunguensis* and *N. gracilis* is due to incomplete lineage sorting in 92 of 149 phylogenies (with topology IDs R2, R4, R9, R11, R15, R16, R17 and R19) and due to introgression in the remaining 57 phylogenies. Under the assumption that the divergence of *N. marunguensis* and *N. gracilis* is younger when the sister group relationship is due to introgression rather than incomplete lineage sorting, we considered the 57 phylogenies with the youngest divergences between *N. marunguensis* and *N. gracilis* to reflect introgression from *N. marunguensis* to *N. gracilis*. To account for the potential bias due to rate variation in different genomic regions, we also used the relative node depth (RND; Rosenzweig *et al.* 2016) as a second criterion to identify phylogenies affected by introgression between *N. marunguensis* to *N. gracilis*. However, the set of 57 phylogenies with the lowest RND between *N. marunguensis* to *N. gracilis* was largely overlapping with the set of 57 phylogenies with the youngest divergences between these species (only four phylogenies were not found in both sets). Thus, we assumed that rate variation did not negatively influence our inference, and used the set based on absolute age estimates as subset (iii). All phylogenies that were not included in any of subsets (i)–(iii) were assumed to be unaffected by introgression and included in subset (iv). Subset (iv) was then used to calculate the ages of most recent common ancestors for all clades of the species tree, whereas divergence times in subsets (i)–(iii) were used to calculate the ages of introgression events. Note that the beginning of a reticulation edge may be older than its end if introgression did not proceed directly between the two species, but via unsampled species. We assume a single time point for each introgression event to be a reasonable assumption, given allopatry in Princess cichlids and the close proximity of bathymetric lines in the localities where the samples used here originate from (except for *N. pulcher*, which originates from the gentle Zambian coastal slopes) (Scholz *et al.* 2007). Proximity of bathymetric lines indicates that the coastline did not have dramatic changes in geographical coordinates during water-level fluctuations throughout the lake's history, which would minimize range changes and secondary contacts (Sturmbauer *et al.* 2001). Nevertheless, the possibility of multiple ancient introgression events involving the same set of species cannot be completely eliminated, a test of which would require the use of multiple populations per species.

Genome-wide landscape of recombination and introgression

Because the linkage map of *Neolamprologus* is not yet known, we indirectly assessed its properties by looking at other molecular parameters that are known to directly correlate with recombination density. Assuming that recombination landscapes are conserved across *Neolamprologus* species and over the timescale of their diversification, a joint local increase in nucleotide diversity, GC content and sequence divergence across the five genomes would likely reflect an increase in local recombination rates (Kulathinal *et al.* 2008; Webster & Hurst 2012; Cutter & Payseur 2013; Roesti *et al.* 2013; Kawakami *et al.* 2014). In addition, we can use SAGUARO breakpoints between segments supporting alternative local topologies to learn about the distribution of the *effective* breaks (*sensu* Hanson 1959) that occurred between genomic regions with different evolutionary histories. We used the sum of branch lengths as a proxy for sequence divergence, unrooted topology U7 for introgression between *N. marunguensis* and the ancestor of *N. pulcher* and *N. olivaceous*, unrooted topology U12 for introgression between *N. brichardi* and *N. pulcher*, unrooted topology U13 as the species tree and all other unrooted topologies as trees affected by incomplete lineage sorting (see Fig. 1 and Table S6, Supporting information). Given that recombination breaks down linkage disequilibrium between loci that prevent interspecific gene flow, any departures from uniform introgression of segments or the number of bases would be informative about the ratio between selection and recombination in hybrids and the progress towards speciation (Barton & Bengtsson 1986; Baird 1995).

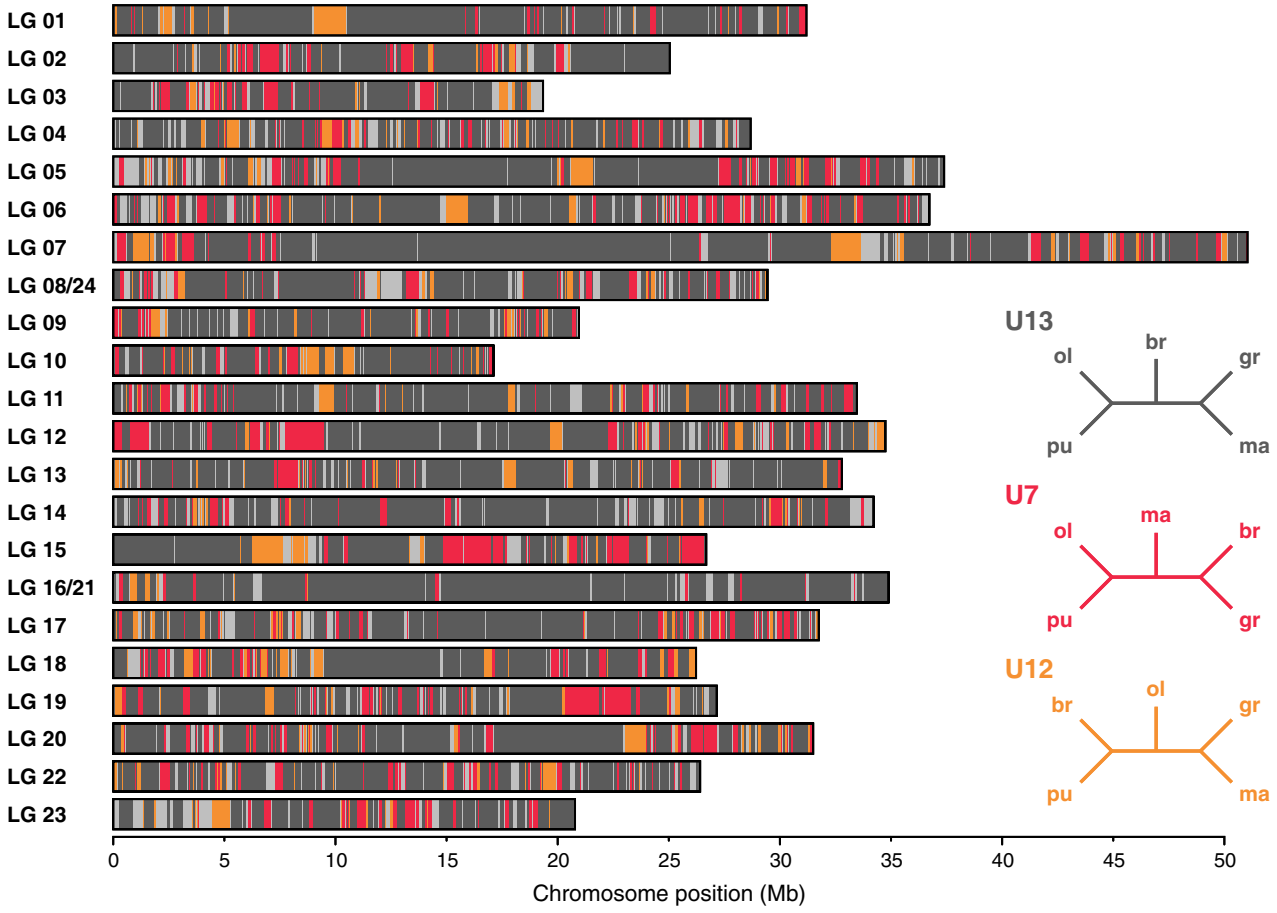

Fig. 1 Distribution of chromosomal segments supporting the three most frequent unrooted local topologies. See Table 1 for further details. Numbers given next to phylogenies indicate topology IDs, as listed in Table S6. Unrooted topologies 7 and 12 support introgression from *N. marunguensis* into the common ancestor of *N. pulcher* and *N. olivaceous* and from *N. brichardi* into *N. pulcher*, respectively. Note that the unrooted topology 13 includes both the most likely species tree and rooted topology 2 (see Table S4). br, *N. brichardi*; gr, *N. gracilis*; ma, *N. marunguensis*; ol, *N. olivaceous*; pu, *N. pulcher*.

Results

Segments supporting local topologies along linkage groups

SAGUARO identified a total of 37 unique similarity matrices and 4781 genomic segments with a length between 21 and 10 613 497 bp (median 46 563 bp). For both the full set of segments and the subset of 224 'most reliable segments', we mapped occurrences of the three most frequently observed unrooted ML topologies of the five *Neolamprologus* species inferred in RAxML along the 23 linkage groups (Fig. 1, Table S6 and Figure S3, Supporting information). A single unrooted topology, in which *N. gracilis* appears as the sister to *N. marunguensis*, and *N. olivaceous* forms a clade with *N. pulcher*, was supported by nearly half (42.9%) of all segments and by the vast majority of the 'most reliable segments' (89.3%). Segments supporting this topology are significantly longer (*t*-test, $P < 10^{-15}$) than all other segments and include the longest overall segment with a length of 10 613 497 bp on linkage group 7. Of all alignment sites, 71.9% are found in segments supporting this topology, and among the 'most reliable segments', this topology is supported by 91.1% of included sites.

Time-calibrated local phylogenies

According to AICM, the strict molecular clock model with rate variation across sites (GTR+Gamma) was the best-fitting model for BEAST analyses of all 426 alignment blocks of 500 000 bp sampled from the 'most reliable segments', followed by the relaxed-clock model without rate variation across sites. For this reason, we here report the results for this model only while presenting more information about results with the other two models in Table S4 (Supporting information). Stationarity of MCMC was indicated by ESS values above 500 for all model parameters, and for all but four of the 426 alignment blocks (less than 1%), the two replicate analyses with different starting positions had converged on the same posterior distribution. The 426 time-calibrated MCC phylogenies represented 18 distinct rooted topologies, all of which are extremely well supported by mean Bayesian posterior probabilities (BPP) (mean BPP across all nodes in all phylogenies: 0.998; Table S4, Supporting information). Per-phylogeny mean age estimates for the first divergence among the five *Neolamprologus* species ranged from 1.67 to 7.28 Ma (median: 3.32; 2.5% quantile: 2.23; 97.5% quantile: 6.23). *Neolamprologus marunguensis* appeared as the sister to the four other *Neolamprologus* in 178 phylogenies, while *N. gracilis* and *N. brichardi* were found in the same position in 92 and 32 phylogenies, respectively. Mean ages of the most recent common ancestor of the five *Neolamprologus* appeared older in phylogenies in which *N. marunguensis* was the first taxon to branch off (3.85 Ma), compared to phylogenies in which *N. gracilis* or *N. brichardi* were the first to diverge (3.71 and 3.22 Ma, respectively).

In the most frequently observed rooted topology, which is found in 129 phylogenies, *N. olivaceous* and *N. pulcher* are sister lineages and form a clade with *N. brichardi*, *N. gracilis* is the sister lineage to this clade, and *N. marunguensis* is the sister species to all of them. The same relationships among the first four species are also supported by the second- and third-most frequent topologies, found in 107 and 67 phylogenies, respectively. Thus, the three most common topologies only differ in the relationship of *N. gracilis* and *N. marunguensis*, which are sisters in 107 phylogenies, while *N. gracilis* is closer to the clade containing *N. olivaceous*, *N. pulcher* and *N. brichardi* in 129 phylogenies, but more distant than *N. marunguensis* in 67 phylogenies.

Local and genome-wide measures of topological incongruence

We used the frequencies of local topologies in three-taxon comparisons to infer the relative contributions of incomplete lineage sorting and introgression to the observed incongruence (Table S5 and Figure S2, Supporting information). *Neolamprologus olivaceous* and *N. pulcher* form a clade in 385 of 426 MCC phylogenies (90.4%) in the comparison of *N. olivaceous*, *N. pulcher* and *N. brichardi*, suggesting that this topology reflects the species tree. In addition, a sister relationship between *N. pulcher* and *N. brichardi* is found in 41 MCC phylogenies, but not a single phylogeny supports *N. olivaceous* and *N. brichardi* to be sisters. This asymmetry between the frequencies of the two alternative topologies is highly unlikely to result from incomplete lineage sorting and thus suggests introgression between *N. pulcher* and *N. brichardi*. Similarly, a clade combining *N. brichardi* and either *N. olivaceous* or *N. pulcher*, relative to *N. marunguensis*, is found in 374 (88%) and 379 three-taxon comparisons (89%), respectively, which thus likely reflects the true species tree. The alternative topologies, however, are found with very different frequencies: *N. marunguensis* appear as sister to either *N. olivaceous* or *N. pulcher* in 50 and 45 phylogenies, respectively, but as the sister of *N. brichardi* in only two phylogenies. As this asymmetry is almost identical regardless of whether *N. olivaceous* or *N. pulcher* are used in the comparison, it suggests ancient introgression between *N. marunguensis* and the common ancestor of *N. olivaceous* and *N. pulcher*. Finally, in the comparison of *N. brichardi*, *N. gracilis* and *N. marunguensis*, all

three possible topologies are found with high frequency, with 185 (43%) phylogenies supporting the most frequent topology in which *N. brichardi* and *N. gracilis* are sisters. The alternative sister group relationships of *N. marunguensis* with either *N. brichardi* or *N. gracilis* are found in 92 (22%) and 149 (35%) phylogenies, respectively. Assuming that the most frequent topology represents the true species tree, differences between these frequencies may be explained by (i) incomplete lineage sorting stochastically leading to minor asymmetry of alternative topologies, (ii) incomplete lineage sorting followed by introgression between *N. marunguensis* and *N. gracilis* or (iii) independent introgression between *N. marunguensis* and both *N. brichardi* and *N. gracilis*.

Using genome-wide tests of excess of shared variants (commonly known as ABBA-BABA tests), we find considerable support for introgression as the most likely explanation for the observed patterns of allele sharing. Considering absolute *z*-scores greater than 3 to be significant (Freedman *et al.* 2014), the null hypothesis of strict bifurcation could be rejected for all but one of the four-taxon comparisons (Table S7, Supporting information). Three comparisons produce particularly strong support for introgression, with *D* statistics < -0.15 and absolute *z*-scores around 60: *Neolamprologus brichardi* shared substantially more variants with *N. pulcher* than with *N. olivaceous* (241 165 *vs.* 152 641; *D*: −0.2248; *z*: −61.99) and *N. marunguensis* shares a similar excess of variants with both *N. pulcher* (279 544 *vs.* 204 089; *D*: −0.1560; *z*: −59.63) and *N. olivaceous* (340 115 *vs.* 246 644; *D*: −0.1593; *z*: −59.62), compared to *N. brichardi*. These test results are consistent with our interpretation of topology frequencies in local phylogenies and further support introgression between *N. pulcher* and *N. brichardi* and between *N. marunguensis* and the common ancestor of both *N. olivaceous* and *N. pulcher*. In addition, *N. marunguensis* shares substantially more variation with *N. gracilis* than with *N. brichardi* (335 990 *vs.* 276 474; *D*: 0.0972; *z*: 27.01). This excess mirrors the asymmetry observed in frequencies of rooted topologies for these three taxa and supports a third case of introgression between *N. marunguensis* and *N. gracilis*.

Maximum-likelihood tests of introgression

Using PhyloNet, we found that the addition of a reticulation edge allowing introgression from *N. brichardi* to *N. pulcher* improves the likelihood by 26.83 log units compared to the introgression-free species tree. About 9% of the genome of *N. pulcher* is estimated to have originated from *N. brichardi* via introgression along this reticulation edge. Adding a second reticulation edge between *N. marunguensis* and the common ancestor of *N. olivaceous* and *N. pulcher* further improves the likelihood by another 14.01 log units, and circa 15% of the genome of this ancestor is estimated to have originated from introgression. Finally, a third reticulation edge allowing for introgression from *N. marunguensis* to *N. gracilis* additionally improves the likelihood by 16.64 log units, and over 25% of the *N. gracilis* genome is estimated to have derived from *N. marunguensis* (Table S8, Supporting information). Taken together, at least three cases of introgression are strongly supported between the five *Neolamprologus* species or their ancestors.

Timing of species divergences and introgression events

Based on the above considerations about inferred introgression events, we used different subsets of trees to calculate the time-calibrated species tree and the timings of introgression (Fig. 2). Diversification of the group is recent and occurred most likely during the Pliocene and Pleistocene, but not earlier than the Late Miocene. Speciation at the base of the tree seems to have been particularly fast, with *N. marunguensis* and *N. gracilis* diverging from their common ancestors almost simultaneously. Speciation of *N. brichardi*, *N. olivaceous* and *N. pulcher* succeeded approximately 1 Ma apart. The three inferred

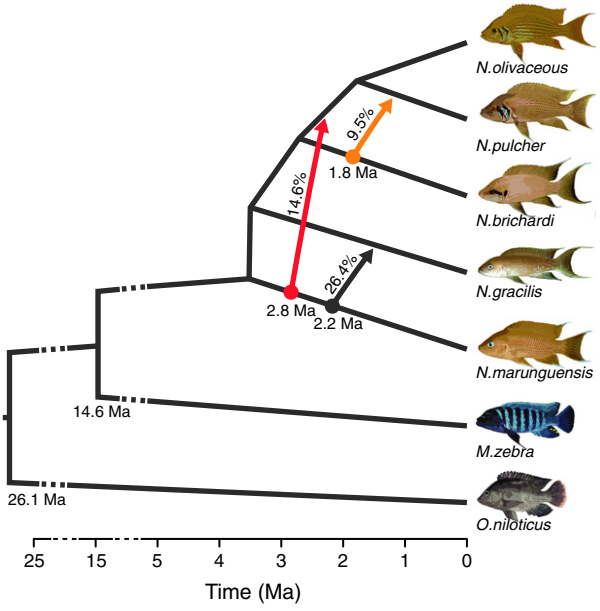

Fig. 2 Time-calibrated species tree of the five *Neolamprologus* and outgroup species inferred with BEAST. Arrows indicate the three most supported introgression events, their mean ages and proportions of introgressed genome. Diversification in this group of *Neolamprologus* took place in the Pliocene and Pleistocene, during the fully lacustrine conditions of Lake Tanganyika. Introgression occurred from 'older' into 'younger' species, possibly indicating a creative role for interspecific gene flow.

introgression events probably occurred at different times (Table S9, Supporting information).

Genome-wide landscape of recombination and introgression

We found small but significant effects of chromosomal position on nucleotide diversity, sequence divergence and GC content (Fig. 3). Nucleotide diversity ($t = 11.34$; $P < 2.2 \times 10^{-16}$; adjusted $R = 0.161$), sequence divergence ($t = 7.68$; $P = 1.9 \times 10^{-14}$; adjusted $R = 0.110$) and GC content ($t = 8.51$; $P < 2.2 \times 10^{-16}$; adjusted $R = 0.121$) all increase with distance from the chromosome centre. Nucleotide diversity and sequence divergence are also positively correlated ($t = 34.52$; $P < 2.2 \times 10^{-16}$; adjusted $R = 0.447$, not shown). Increases in nucleotide diversity, sequence divergence and GC content are known effects of increased local recombination (Webster & Hurst 2012), which indicates that the recombination landscape is not uniform along the chromosomes of (*Neolamprologus*) cichlids and that its density increases towards chromosomal ends. Using breakpoint information and segments that support different local topologies from SAGUARO, we found increasing number of segments away from the chromosome centre, indicating an increase in the number of *effective* recombination events (Kolmogorov–Smirnov two-sided test for uniform distribution of segments: $D = 0.11574$; $P < 2.2 \times 10^{-16}$; segment size: $t = -7.11$; $P < 1.4 \times 10^{-12}$; adjusted $R = 0.101$). Finally, we found that introgression is not uniform along the chromosomes (Fig. 4). In particular, introgressed segments are more abundant towards the periphery of linkage groups and not uniformly scattered around the genome (introgression from *N. marunguensis*: Kolmogorov–Smirnov two-sided test: $D = 0.12$; $P = 6.5 \times 10^{-11}$; introgression from *N. brichardi*: $D = 0.16$; $P = 4.7 \times 10^{-8}$). The proportion of introgressed bases also increases away from the chromosome centre (introgression from *N. marunguensis*: $t = 3.99$; $P = 1.29 \times 10^{-4}$; adjusted $R = 0.362$; introgression from *N. brichardi*: $t = 1.69$; $P = 0.094$; adjusted $R = 0.136$), as does the proportion of bases supporting incomplete lineage sorting and other phylogenetic noise ($t = 3.20$; $P = 1.85 \times 10^{-3}$; adjusted $R = 0.292$). In the opposite direction, the proportion of bases supporting the correct species tree decreases with distance from the chromosome centre ($t = -5.67$; $P = 1.4 \times 10^{-7}$; adjusted $R = 0.489$).

Discussion

The Neolamprologus species tree, incomplete lineage sorting and introgressive hybridization

Here, we illuminate the complex evolutionary history of a highly diverse group of East African cichlid fishes that were previously shown to be involved in introgressive hybridisation (Salzburger *et al.* 2002; Gante & Salzburger 2012). By employing a combination of whole-genome sequencing and phylogenomic analyses of five closely related *Neolamprologus* species, we determined the most likely species tree, found concordant evidence for incomplete lineage sorting or interspecific gene flow affecting different branches of the phylogeny and determined the direction and timing of introgression events involving these species.

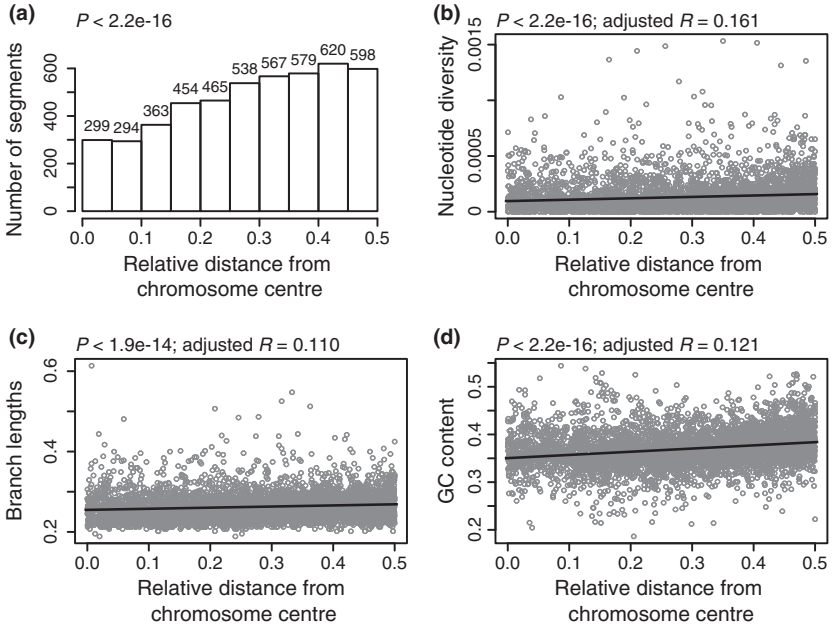

Fig. 3 Bias in the genomic distribution of molecular parameters, all segment topologies considered. In (a), the number of segments increases with distance from chromosome centre; that is, the number of breakpoints between segments increases, indicating that recombination rate is lowest closer to the chromosome centre. In (b), (c) and (d), nucleotide diversity, branch lengths (used as a proxy for sequence divergence) and GC content, respectively, all increase with distance from chromosome centre. These parameters are known to correlate with recombination rate, supporting the use of breakpoint data to infer the recombination landscape.

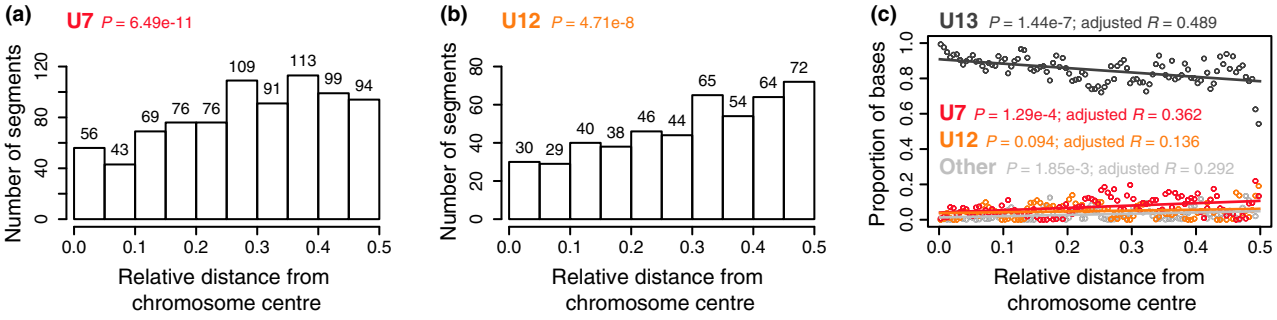

Fig. 4 Bias in the genomic distribution of introgression and incomplete lineage sorting. In (a), the number of segments with unrooted topology 7 (introgression from *N. marunguensis* to the ancestor of *N. pulcher* and *N. olivaceous*) and (b) the number of segments with unrooted topology 12 (introgression between *N. brichardi* and *N. pulcher*) are lowest closest to the chromosome centre and increase towards the periphery. In (c), we show the linear relation between the proportion of bases supporting different topologies and chromosome position. The proportion of bases supporting introgression (U7 and U12) also increases with distance from chromosome centre, as does the proportion of bases affected by incomplete lineage sorting and other phylogenetic noise ('other' in light grey). As a consequence, the proportion of bases supporting the species tree (U13) decreases with increasing distance from chromosome centre. In Figure S4 (Supporting information), we show the same data as in (c) fitted with polynomial functions. The results are qualitatively the same, except that the more complex polynomial functions are better in describing the relationship between distance from chromosome centre and the proportion of bases supporting alternative topologies.

While it has been shown that with high levels of introgression, the most frequently observed topology can differ from the actual species tree (e.g. Fontaine *et al.* 2015), we here conclude that this is not the case for the five *Neolamprologus* species, based on several lines of evidence. First, the most common topology can usually be considered the most probable species tree topology unless additional evidence supports a different scenario. Second, the unrooted topology (U13) corresponding to the most frequent rooted topology (R1) is predominantly found in low-recombination regions in the centre of chromosomes (Figs 1 and 4), which are known to support the true species tree topology more reliably than regions with high recombination (Pease & Hahn 2013). Third, the age estimate for the divergence of *Neolamprologus* is older for trees with the most frequent rooted topology (R1), compared to trees with other frequently occurring topologies (R2, R3; Table S4, Supporting information) in which divergence age estimates may have been reduced due to introgression. Thus, the most likely species tree is concordant with the most frequently observed rooted tree topology (R1), in which *N. olivaceous* and *N. pulcher* are sister lineages and form a clade with *N. brichardi*, with *N. gracilis* being their sister lineage and *N. marunguensis* being the sister species to all of them. This species tree suggests that the organized facial pigmentation, characteristic of many of the 'Princess cichlids' (Bachmann *et al.* 2016), evolved within this group of pigmented species from nonpigmented cooperatively breeding ancestors (Fig. 2).

As expected for rapidly diversifying taxa, we found evidence for incomplete lineage sorting affecting short internodes in the species tree (Pease & Hahn 2013), in particular at the base of *Neolamprologus* and involving *N. brichardi*, *N. gracilis* and *N. marunguensis*. Patterns of allele sharing and marker incongruence are known to affect the phylogenies of diverse groups, cichlids in particular (Rüber *et al.* 2001; Koblmüller *et al.* 2007, 2010; Joyce *et al.* 2011; Genner & Turner 2012; Schwarzer *et al.* 2012; Keller *et al.* 2013; Wagner *et al.* 2013; Brawand *et al.* 2014; Ford *et al.* 2015; Martin *et al.* 2015a; Meyer *et al.* 2015). Nevertheless, we were able to disentangle the effects of incomplete lineage sorting from those of introgression. This allowed us to test the hypothesis of hybrid origin of *N. marunguensis* (Salzburger *et al.* 2002) and to explore the causes for polyphyly in the relationship between *N. brichardi* and *N. pulcher* (Duftner *et al.* 2007).

Incongruent patterns of mitochondrial and microsatellite allele sharing among *N. marunguensis*, *N. gracilis* and *N. olivaceous* suggested that the first could be of hybrid origin (Salzburger *et al.* 2002), although multilocus nuclear AFLP (amplified fragment length polymorphism, a restriction enzyme-based method) supported their reciprocal monophyly (Sturmbauer *et al.* 2010). While we confirm episodes of introgressive hybridization involving these three species, in proportions similar to those observed in *Heliconius* butterflies (Martin *et al.* 2013), our results indicate an alternative scenario of introgression: based on whole genomes, we show that *N. marunguensis* was the donor of substantial proportions of genetic material to *N. gracilis* and to *N. olivaceous* (and *N. pulcher*), rather than being a recipient, hybrid species. Thus, we find no support for the hybrid origin hypothesis of *N. marunguensis* (Salzburger *et al.* 2002), which is not surprising given

the extreme difficulty in robustly distinguishing hybrid speciation from introgressive hybridization, particularly when the original hypotheses are based on a few neutral genetic markers without reference to adaptive hybrid traits (Mallet 2007; Jiggins *et al.* 2008). Our findings using whole genomes enlarge the pool of genome-wide studies that reject previous hybrid origin hypotheses in animals (e.g. Meyer *et al.* 2006; Schumer *et al.* 2013), even in systems where hybrid trait speciation has found strong support, like in *Heliconius* butterflies (Mavárez *et al.* 2006; Kronforst *et al.* 2007; Salazar *et al.* 2010). This current tendency of rejecting possible instances of hybrid speciation based on genome-wide data raises the possibility that homoploid hybrid speciation may in fact be exceptionally rare in animals.

In addition, we find strong evidence for nuclear introgression from *N. brichardi* into *N. pulcher*. These species were previously found to be polyphyletic based on mitochondrial DNA sequences, but reciprocally monophyletic based on AFLP data (Duftner *et al.* 2007; Sturmbauer *et al.* 2010). In addition, while *N. olivaceous* and *N. pulcher* were consistently placed into divergent mitochondrial lineages within the lamprologines (Salzburger *et al.* 2002; Day *et al.* 2007; Duftner *et al.* 2007; Sturmbauer *et al.* 2010), a sister species relationship is strongly supported by our genome-wide analyses. In the light of these results, we reinterpret the previously reported mitochondrial polyphyly as resulting from widespread mitochondrial introgression rather than a lack of distinctiveness between *N. brichardi* and *N. pulcher* (Duftner *et al.* 2007). The distant relationship between *N. olivaceous* and other *Neolamprologus* (namely *N. pulcher*) inferred from mtDNA also likely reflects ancient mitochondrial introgression. Altogether, these results point towards a bias in nuclear *vs.* mitochondrial introgression (Chan & Levin 2005; Carson & Dowling 2006; Good *et al.* 2015; Patten *et al.* 2015): *N. marunguensis* acts as a major donor of nuclear genes, while it is also a receiver of mitochondrial genomes, at least in some populations. Whether alternative mitochondrial lineages have become fixed in different allopatric populations, or have been entirely replaced in some species, is presently unknown but is a plausible scenario, particularly in *N. pulcher* and *N. olivaceous* (see Nevado *et al.* 2009 for another Lamprologini example).

Speciation and introgression in Lamprologini and other cichlids

The tribe Lamprologini, to which *Neolamprologus* belongs, is the most species rich and also one of the most phenotypically diverse (in morphology, behaviour, life history) lineages of cichlids in Lake Tanganyika (Stiassny 1997; Gante & Salzburger 2012). We are just starting to acknowledge the various trophic adaptations (Muschick *et al.* 2012) and behaviours (Heg & Bachar 2006) that likely contributed to the lineage's evolutionary success. In addition, their genomes are very dynamic compared to Nile tilapia and other teleost genomes, with new gene duplications, accelerated coding sequence evolution, an abundance of noncoding element divergence, expression divergence linked to transposable element insertions, regulation by novel microRNAs (Brawand *et al.* 2014) and extensive introgressive hybridization as shown here and elsewhere (Salzburger *et al.* 2002; Schelly *et al.* 2006; Day *et al.* 2007; Koblmüller *et al.* 2007, 2010; Nevado *et al.* 2009; Sturmbauer *et al.* 2010). While the proximate consequences of introgression remain unknown in lamprologines, it seems to have occurred *throughout* their evolutionary history, in contrast with cichlids from other African Great Lakes where introgression has been hypothesized to fuel the *onset* of adaptive radiations (Seehausen 2004).

Given that many lamprologine species have deme-like allopatric distributions, frequently fragmented and interspersed by other species, the alternation between periods of sympatry and allopatry could be driving introgression and the evolution of the lamprologines' genomes as we know them. During periods in allopatry, differences should accumulate in chromosome centres, while gene flow should resume to a greater extent in chromosome ends in periods of sympatry, recombination and selection permitting. Lake-level fluctuations, which have happened multiple times during the complex geological and palaeo-climatological history of Lake Tanganyika (Cohen *et al.* 1997; Salzburger *et al.* 2014), are potential drivers of changes in species ranges (Sturmbauer *et al.* 2001). Therefore, it is possible that the patterns of gene flow observed in *Neolamprologus* have involved species or populations in addition to the ones studied here and that different demes of each species tell a different story of gene flow and incomplete lineage sorting. It is also very likely that introgressive hybridization has also occurred at higher taxonomic levels, such as between tribes, or at the genesis of hybrid tribes, which would explain the difficulties in generating well-supported nuclear phylogenies of the cichlid assemblage from Lake Tanganyika (Meyer *et al.* 2015, 2016). Another fascinating aspect of the cichlid model systems is the emergence of parallel phenotypes in different cichlid lineages, as well as convergence at the molecular level (Salzburger 2009). Whole-genome data will allow testing whether these were generated *de novo* or *recycled* from pre-existing variation through introgression, transferring 'recycled genetic pathways and phenotypes' across geographical scales, even between lakes thought to be watertight (Jiggins 2014).

Recombination bias, linked selection and introgression

We found significant correlations between chromosomal position and several molecular parameters, such as nucleotide diversity, sequence divergence and GC content, all of which increase towards the chromosome periphery, likely modulated by recombination rate as seen in a number of other systems (Begun & Aquadro 1992; Hellmann *et al.* 2003; Begun *et al.* 2007; Kulathinal *et al.* 2008; Sella *et al.* 2009; McGaugh *et al.* 2012; Roesti *et al.* 2013; Campos *et al.* 2014; Good *et al.* 2014; Kawakami *et al.* 2014; Tine *et al.* 2014; Burri *et al.* 2015). An increase in recombination rate with increased distance from chromosome centres is in line with observations from pedigree studies (Jensen-Seaman *et al.* 2004; Backström *et al.* 2010; Niehuis *et al.* 2010; Brunschwig *et al.* 2012; Tortereau *et al.* 2012; Roesti *et al.* 2013; Kawakami *et al.* 2014; Burri *et al.* 2015; Phillips *et al.* 2015). Because cichlids have mostly acrocentric, subtelocentric and telocentric chromosomes (Mazzuchelli *et al.* 2012), the observed patterns of variation from centre to periphery are likely not driven by centromere position.

In addition, SAGUARO breakpoint density also increases towards the end of chromosomes in *Neolamprologus*, indicating that *effective* recombination events (between stretches of DNA with different evolutionary histories, *sensu* Hanson 1959) are not uniformly distributed along the chromosomes. In fact, we observe increased introgression towards the chromosome periphery, which suggests that selection is shaping the genomic architecture as a function of recombination rate. Cline theory predicts that recombination rate influences the magnitude of selection at linked loci: barriers to gene flow are more effective in regions of low recombination, while increased recombination makes selection more efficient at individual loci, thereby reducing Hill–Robertson interference and allowing the introgression of neutral or positively selected loci (Haldane 1948; Hill & Robertson 1966; Barton 1983; Barton & Hewitt 1985; Barton & Bengtsson 1986; Baird 1995; Gavrilets 2004; Charlesworth 2009; Bierne *et al.* 2011; Flaxman *et al.* 2014; Roesti *et al.* 2014). This is so because larger segments are more likely to harbour alleles that are deleterious in a hybrid background and are removed by selection before recombination can occur, and thus, *effective* introgression is more likely in regions where linkage is more rapidly broken down by recombination (Barton 1983; Barton & Hewitt 1985; Barton & Bengtsson 1986; Baird 1995; Martinsen *et al.* 2001). Still in the absence of recombination biases, theory predicts that the strength of a barrier to gene flow is stronger at the chromosome centre than at its periphery (Barton & Bengtsson 1986), even if we observe an increase in nucleotide diversity, sequence divergence and GC content suggestive of some degree of recombination bias.

Reduced introgression has been predicted and indeed observed in regions of low recombination, such as inversions and around centromeres, which have long been thought to contribute to the accumulation of genetic differences between species (e.g. Rieseberg *et al.* 1999; Butlin 2005; Turner *et al.* 2005; Kirkpatrick & Barton 2006; Yatabe *et al.* 2007; Kulathinal *et al.* 2009; Carneiro *et al.* 2010; Geraldes *et al.* 2011; Ellegren *et al.* 2012; Nachman & Payseur 2012; Lohse *et al.* 2015; Roesti *et al.* 2015), although high differentiation could also be driven by linked selection alone (Burri *et al.* 2015). Here, we show that *collinear* regions with low recombination are also more resistant to introgression. Hence, isolation (speciation) genes can accumulate in different lowly recombining regions without the restrictive precondition of close physical proximity, generating unlinked islands of differentiation. Therefore, these regions simultaneously show reduced gene flow and reduced diversity. Our results support a scenario in which 'genomic islands of differentiation' behave simultaneously as regions of low recombination and low introgression, resolving an artificial dichotomy between potentially complementary processes (Cruickshank & Hahn 2014). Effective introgression of small-sized segments broken down by repeated recombination is then more likely with increasing distance from the chromosome centre. Peripheral regions experiencing higher recombination and more effective selection regimes thus show increased evolvability and are likely better at responding to changes in spatially or temporally heterogeneous environments (Tigano & Friesen 2016). Conversely, chromosome centres with lower recombination experience higher interference, increased coupling and lower introgression. Linkage disequilibrium can thus be created and maintained without the need for tight physical linkage, perhaps in structured populations or after a period of allopatry. Intrinsic Bateson–Dobzhansky–Muller incompatibilities fit right in this scenario (Coyne & Orr 2004; Gavrilets 2004).

Implications of heterogeneous recombination landscapes for speciation in Neolamprologus and other systems

The recurrent emergence of genomic regions that simultaneously show high differentiation (reduced gene flow) and low genetic diversity across taxa suggests that shared genomic features contribute to the build-up of genomic islands of differentiation and speciation (Chowdhury *et al.* 2009; Kulathinal *et al.* 2009; Rockman & Kruglyak 2009; Backström *et al.* 2010; Bradley *et al.* 2011; Roesti *et al.* 2013; Kawakami *et al.* 2014; Burri *et al.* 2015). Indeed, heterogeneous yet collinear genomes are

generated by biases in recombination rate across the chromosome (chromosome centre-biased divergence, *sensu* Roesti *et al.* 2012). These biases generate slower-evolving, less-diverse and introgression-resistant chromosome centres that can act as 'centres of incompatibility', while the faster-evolving, more-diverse and more mutation- and gene flow-prone chromosome ends are 'peripheries of evolvability', that is regions where molecular adaptations can originate at a faster pace and spread across the population. Based on these findings, we anticipate that genomic landscapes modulated by recombination and selection relate to observed levels of taxonomic and phenotypic diversity in cichlids and other taxa with collinear genomes, linking genomic architecture to observed phenotypic diversity. While more complex and realistic models and simulations are already being developed (e.g. Flaxman *et al.* 2012, 2013, 2014; Roesti *et al.* 2014), promising approaches should incorporate and explore the pervasive effects of heterogeneous genomic landscapes and selection at linked sites on speciation, adaptation and gene flow.

The fact that we observe this dichotomy between chromosome centres and periphery in *Neolamprologus* could explain, from a genomic perspective, both the high speciation rate and extreme adaptability observed in African cichlid radiations. The cichlid system offers an ideal opportunity to further test the intrinsic properties of genomes in driving speciation and adaptation, as cichlids are composed of several lineages with various levels of taxonomic and phenotypic diversity (Gante & Salzburger 2012).

Acknowledgements

We are indebted to Heinz Buescher for providing Congolese samples of *Neolamprologus* and thank Marius Roesti and Daniel Berner for valuable discussions. We acknowledge the Norwegian Sequencing Centre (NSC; http://www.sequencing.uio.no), the computational resources of the Abel Cluster at the University of Oslo and the Norwegian metacenter for High Performance Computing (NOTUR). This work was supported by 'University of Basel Excellence Scholarships for Young Researchers' and 'Novartis Excellence Scholarships for Life Sciences' to HFG, and the European Research Council (CoG 'CICHLID~X') and the Swiss National Science Foundation (SNF) to WS.

References

Abbott R, Albach D, Ansell S *et al.* (2013) Hybridization and speciation. *Journal of Evolutionary Biology*, **26**, 229–246.

Anderson E (1949) *Introgressive Hybridization*. Wiley, New York.

Arnold ML (2006) *Evolution Through Genetic Exchange*. Oxford University Press, Oxford.

Azuma Y, Kumazawa Y, Miya M, Mabuchi K, Nishida M (2008) Mitogenomic evaluation of the historical biogeography of cichlids toward reliable dating of teleostean divergences. *BMC Evolutionary Biology*, **8**, 215.

Baack EJ, Rieseberg LH (2007) A genomic view of introgression and hybrid speciation. *Current Opinion in Genetics and Development*, **17**, 513–518.

Bachmann JC, Cortesi F, Hall M *et al.* (2016) Social selection maintains honesty of a dynamic visual signal in cichlid fish. *bioRxiv*, 1–19, doi: 10.1101/039552.

Backström N, Forstmeier W, Schielzeth H *et al.* (2010) The recombination landscape of the zebra finch *Taeniopygia guttata* genome. *Genome Research*, **20**, 485–495.

Baele G, Lemey P, Bedford T *et al.* (2012) Improving the accuracy of demographic and molecular clock model comparison while accommodating phylogenetic uncertainty. *Molecular Biology and Evolution*, **29**, 2157–2167.

Baird SJE (1995) A simulation study of multilocus clines. *Evolution*, **49**, 1038–1045.

Balshine S, Leach B, Neat F *et al.* (2001) Correlates of group size in a cooperatively breeding cichlid fish (*Neolamprologus pulcher*). *Behavioral Ecology and Sociobiology*, **50**, 134–140.

Barton NH (1983) Multilocus clines. *Evolution*, **37**, 454–471.

Barton N, Bengtsson BO (1986) The barrier to genetic exchange between hybridising populations. *Heredity*, **57**, 357–376.

Barton NH, Hewitt GM (1985) Analysis of hybrid zones. *Annual Review of Ecology and Systematics*, **16**, 113–148.

Begun DJ, Aquadro CF (1992) Levels of naturally occurring DNA polymorphism correlate with recombination rates in *D. melanogaster*. *Nature*, **356**, 519–520.

Begun DJ, Holloway AK, Stevens K *et al.* (2007) Population genomics: whole-genome analysis of polymorphism and divergence in *Drosophila simulans*. *PLoS Biology*, **5**, e310.

Berg PR, Jentoft S, Star B *et al.* (2015) Adaptation to low salinity promotes genomic divergence in Atlantic cod (*Gadus morhua* L.). *Genome Biology and Evolution*, **7**, 1644–1663.

Berner D, Salzburger W (2015) The genomics of organismal diversification illuminated by adaptive radiations. *Trends in Genetics*, **31**, 491–499.

Bierne N, Welch J, Loire E, Bonhomme F, David P (2011) The coupling hypothesis: why genome scans may fail to map local adaptation genes. *Molecular Ecology*, **20**, 2044–2072.

Bouckaert R, Heled J, Kühnert D *et al.* (2014) BEAST 2: a software platform for Bayesian evolutionary analysis. *PLoS Computational Biology*, **10**, e1003537.

Bradley KM, Breyer JP, Melville DB *et al.* (2011) An SNP-based linkage map for zebrafish reveals sex determination loci. *G3-Genes|Genomes|Genetics*, **1**, 3–9.

Brawand D, Wagner CE, Li YI *et al.* (2014) The genomic substrate for adaptive radiation in African cichlid fish. *Nature*, **513**, 375–381.

Brunschwig H, Levi L, Ben-David E *et al.* (2012) Fine-scale maps of recombination rates and hotspots in the mouse genome. *Genetics*, **191**, 757–764.

Burri R, Nater A, Kawakami T *et al.* (2015) Linked selection and recombination rate variation drive the evolution of the genomic landscape of differentiation across the speciation continuum of Ficedula flycatchers. *Genome Research*, **25**, 1–10.

Büscher HH (1997) Ein neuer Cichlide aus dem Tanganyikasee: *Neolamprologus helianthus* (Cichlidae, Lamprologini). *Die Aquarien- und Terrarien-Zeitschrift (DATZ)*, **50**, 701–706.

Butlin RK (2005) Recombination and speciation. *Molecular Ecology*, **14**, 2621–2635.

Campos JL, Halligan DL, Haddrill PR, Charlesworth B (2014) The relation between recombination rate and patterns of molecular evolution and variation in *Drosophila melanogaster*. *Molecular Biology and Evolution*, **31**, 1010–1028.

Carneiro M, Blanco-Aguiar JA, Villafuerte R, Ferrand N, Nachman MW (2010) Speciation in the European rabbit (*Oryctolagus cuniculus*): islands of differentiation on the X chromosome and autosomes. *Evolution*, **64**, 3443–3460.

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*. *Molecular Ecology*, **15**, 667–679.

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

Charlesworth B (2009) Effective population size and patterns of molecular evolution and variation. *Nature Reviews Genetics*, **10**, 195–205.

Chowdhury R, Bois PRJ, Feingold E, Sherman SL, Cheung VG (2009) Genetic analysis of variation in human meiotic recombination. *PLoS Genetics*, **5**, e1000648.

Cohen AS, Soreghan MJ, Scholz CA (1993) Estimating the age of formation of lakes: an example from Lake Tanganyika, East African Rift system. *Geology*, **21**, 511–514.

Cohen AS, Lezzar K-E, Tiercelin J-J, Soreghan M (1997) New palaeogeographic and lake-level reconstructions of Lake Tanganyika: implications for tectonic, climatic and biological evolution in a rift lake. *Basin Research*, **9**, 107–132.

Coyne JA, Orr HA (2004) *Speciation*. Sinauer Associates Inc., Sunderland, Massachusetts, US.

Cruickshank TE, Hahn MW (2014) Reanalysis suggests that genomic islands of speciation are due to reduced diversity, not reduced gene flow. *Molecular Ecology*, **23**, 3133–3157.

Cui R, Schumer M, Kruesi K *et al.* (2013) Phylogenomics reveals extensive reticulate evolution in Xiphophorus fishes. *Evolution*, **67**, 2166–2179.

Cutter AD, Payseur BA (2013) Genomic signatures of selection at linked sites: unifying the disparity among species. *Nature Reviews Genetics*, **14**, 262–274.

Danecek P, Auton A, Abecasis G *et al.* (2011) The variant call format and VCFtools. *Bioinformatics*, **27**, 2156–2158.

Day JJ, Santini S, Garcia-Moreno J (2007) Phylogenetic relationships of the Lake Tanganyika cichlid tribe Lamprologini: the story from mitochondrial DNA. *Molecular Phylogenetics and Evolution*, **45**, 629–642.

DeMarais BD, Dowling TE, Douglas ME, Minckley WL, Marsh PC (1992) Origin of *Gila seminuda* (Teleostei: Cyprinidae) through introgressive hybridization: implications for evolution and conservation. *Proceedings of the National Academy of Sciences of the United States of America*, **89**, 2747–2751.

Dowling TE, Secor CL (1997) The role of hybridization and introgression in the diversification of animals. *Annual Review of Ecology and Systematics*, **28**, 593–619.

Duftner N, Sefc KM, Koblmüller S *et al.* (2007) Parallel evolution of facial stripe patterns in the *Neolamprologus brichardi/pulcher* species complex endemic to Lake Tanganyika. *Molecular Phylogenetics and Evolution*, **45**, 706–715.

Durand EY, Patterson N, Reich D, Slatkin M (2011) Testing for ancient admixture between closely related populations. *Molecular Biology and Evolution*, **28**, 2239–2252.

Eaton DAR, Hipp AL, González-Rodríguez A, Cavender-Bares J (2015) Historical introgression among the American live oaks and the comparative nature of tests for introgression. *Evolution*, **69**, 2587–2601.

Ellegren H, Smeds L, Burri R *et al.* (2012) The genomic landscape of species divergence in *Ficedula* flycatchers. *Nature*, **491**, 756–760.

Feder JL, Egan SP, Nosil P (2012) The genomics of speciation-with-gene-flow. *Trends in Genetics*, **28**, 342–350.

Felsenstein J (1985) Confidence limits on phylogenies: an approach using the bootstrap. *Evolution*, **39**, 783–791.

Flaxman SM, Feder JL, Nosil P (2012) Spatially explicit models of divergence and genome hitchhiking. *Journal of Evolutionary Biology*, **25**, 2633–2650.

Flaxman SM, Feder JL, Nosil P (2013) Genetic hitchhiking and the dynamic buildup of genomic divergence during speciation with gene flow. *Evolution*, **67**, 2577–2591.

Flaxman SM, Wacholder AC, Feder JL, Nosil P (2014) Theoretical models of the influence of genomic architecture on the dynamics of speciation. *Molecular Ecology*, **23**, 4074–4088.

Fontaine MC, Pease JB, Steele A *et al.* (2015) Extensive introgression in a malaria vector species complex revealed by phylogenomics. *Science*, **347**, 1258524.

Ford AGP, Dasmahapatra KK, Rüber L *et al.* (2015) High levels of interspecific gene flow in an endemic cichlid fish adaptive radiation from an extreme lake environment. *Molecular Ecology*, **24**, 3421–3440.

Freedman AH, Gronau I, Schweizer RM *et al.* (2014) Genome sequencing highlights the dynamic early history of dogs. *PLoS Genetics*, **10**, e1004016.

Friedman M, Keck BP, Dornburg A *et al.* (2013) Molecular and fossil evidence place the origin of cichlid fishes long after Gondwanan rifting. *Proceedings of the Royal Society B: Biological Sciences*, **280**, 20131733.

Gante HF, Salzburger W (2012) Evolution: cichlid models on the runaway to speciation. *Current Biology*, **22**, R956–R958.

Gavrilets S (2004) *Fitness Landscapes and the Origin of Species*. Princeton University Press, Princeton.

Genner MJ, Turner GF (2012) Ancient hybridization and phenotypic novelty within Lake Malawi's cichlid fish radiation. *Molecular Biology and Evolution*, **29**, 195–206.

Geraldes A, Basset P, Smith KL, Nachman MW (2011) Higher differentiation among subspecies of the house mouse (*Mus musculus*) in genomic regions with low recombination. *Molecular Ecology*, **20**, 4722–4736.

Gompert Z, Lucas LK, Buerkle CA *et al.* (2014) Admixture and the organization of genetic diversity in a butterfly species complex revealed through common and rare genetic variants. *Molecular Ecology*, **23**, 4555–4573.

Good BH, Walczak AM, Neher RA, Desai MM (2014) Genetic diversity in the interference selection limit. *PLoS Genetics*, **10**, e1004222.

Good JM, Vanderpool D, Keeble S, Bi K (2015) Negligible nuclear introgression despite complete mitochondrial capture between two species of chipmunks. *Evolution*, **69**, 1961–1972.

Green RE, Krause J, Briggs AW *et al.* (2010) A draft sequence of the Neandertal genome. *Science*, **328**, 710–722.

Haldane JBS (1948) The theory of a cline. *Journal of Genetics*, **48**, 277–284.

Hanson WD (1959) The breakup of initial linkage blocks under selected mating systems. *Genetics*, **44**, 857–868.

Heg D, Bachar Z (2006) Cooperative breeding in the Lake Tanganyika cichlid *Julidochromis ornatus*. *Environmental Biology of Fishes*, **76**, 265–281.

Heg D, Bachar Z, Brouwer L, Taborsky M (2004) Predation risk is an ecological constraint for helper dispersal in a cooperatively breeding cichlid. *Proceedings of the Royal Society B: Biological Sciences*, **271**, 2367–2374.

Heliconius Genome Consortium (2012) Butterfly genome reveals promiscuous exchange of mimicry adaptations among species. *Nature*, **487**, 94–98.

Hellmann I, Ebersberger I, Ptak SE, Pääbo S, Przeworski M (2003) A neutral explanation for the correlation of diversity with recombination rates in humans. *American Journal of Human Genetics*, **72**, 1527–1535.

Hermansen JS, Haas F, Trier CN *et al.* (2014) Hybrid speciation through sorting of parental incompatibilities in Italian sparrows. *Molecular Ecology*, **23**, 5831–5842.

Hill WG, Robertson A (1966) The effect of linkage on limits to artificial selection. *Genetical Research*, **8**, 269–294.

Huerta-Sánchez E, Jin X, Bianba Z *et al.* (2014) Altitude adaptation in Tibetans caused by introgression of Denisovan-like DNA. *Nature*, **512**, 194–197.

Jensen-Seaman MI, Furey TS, Payseur BA *et al.* (2004) Comparative recombination rates in the rat, mouse, and human genomes. *Genome Research*, **14**, 528–538.

Jiggins CD (2014) Evolutionary biology: radiating genomes. *Nature*, **513**, 318–319.

Jiggins CD, Salazar C, Linares M, Mavarez J (2008) Hybrid trait speciation and *Heliconius* butterflies. *Philosophical Transactions of the Royal Society of London Series B, Biological Sciences*, **363**, 3047–3054.

Joyce DA, Lunt DH, Genner MJ *et al.* (2011) Repeated colonization and hybridization in Lake Malawi cichlids. *Current Biology*, **21**, R108–R109.

Kawakami T, Smeds L, Backström N *et al.* (2014) A high-density linkage map enables a second-generation collared flycatcher genome assembly and reveals the patterns of avian recombination rate variation and chromosomal evolution. *Molecular Ecology*, **23**, 4035–4058.

Keller I, Wagner CE, Greuter L *et al.* (2013) Population genomic signatures of divergent adaptation, gene flow and hybrid speciation in the rapid radiation of Lake Victoria cichlid fishes. *Molecular Ecology*, **22**, 2848–2863.

Kirkpatrick M, Barton N (2006) Chromosome inversions, local adaptation and speciation. *Genetics*, **173**, 419–434.

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

Koblmüller S, Egger B, Sturmbauer C, Sefc KM (2010) Rapid radiation, ancient incomplete lineage sorting and ancient hybridization in the endemic Lake Tanganyika cichlid tribe Tropheini. *Molecular Phylogenetics and Evolution*, **55**, 318–334.

Kocher TD (2004) Adaptive evolution and explosive speciation: the cichlid fish model. *Nature Reviews Genetics*, **5**, 288–298.

Konings A (1998) *Tanganyika Cichlids in Their Natural Habitat*. 2nd edn. Cichlid Press, El Paso, Texas.

Kronforst MR, Salazar C, Linares M, Gilbert LE (2007) No genomic mosaicism in a putative hybrid butterfly species. *Proceedings of the Royal Society B: Biological Sciences*, **274**, 1255–1264.

Kulathinal RJ, Bennett SM, Fitzpatrick CL, Noor MAF (2008) Fine-scale mapping of recombination rate in *Drosophila* refines its correlation to diversity and divergence. *Proceedings of the National Academy of Sciences of the United States of America*, **105**, 10051–10056.

Kulathinal RJ, Stevison LS, Noor MAF (2009) The genomics of speciation in *Drosophila*: diversity, divergence, and introgression estimated using low-coverage genome sequencing. *PLoS Genetics*, **5**, e1000550.

Lamichhaney S, Berglund J, Almén MS *et al.* (2015) Evolution of Darwin's finches and their beaks revealed by genome sequencing. *Nature*, **518**, 371–375.

Li H (2011) A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*, **27**, 2987–2993.

Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*, **25**, 1754–1760.

Li H, Handsaker B, Wysoker A *et al.* (2009) The sequence alignment/map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.

Liu KJ, Steinberg E, Yozzo A *et al.* (2015) Interspecific introgressive origin of genomic diversity in the house mouse. *Proceedings of the National Academy of Sciences*, **112**, 196–201.

Lohse K, Clarke M, Ritchie MG, Etges WJ (2015) Genome-wide tests for introgression between cactophilic *Drosophila* implicate a role of inversions during speciation. *Evolution*, **69**, 1178–1190.

Malinsky M, Challis RJ, Tyers AM *et al.* (2015) Genomic islands of speciation separate cichlid ecomorphs in an East African crater lake. *Science*, **350**, 1493–1498.

Mallet J (2007) Hybrid speciation. *Nature*, **446**, 279–283.

Martin SH, Dasmahapatra KK, Nadeau NJ *et al.* (2013) Genome-wide evidence for speciation with gene flow in *Heliconius* butterflies. *Genome Research*, **23**, 1817–1828.

Martin CH, Cutler JS, Friel JP *et al.* (2015a) Complex histories of repeated gene flow in Cameroon crater lake cichlids cast doubt on one of the clearest examples of sympatric speciation. *Evolution*, **69**, 1406–1422.

Martin SH, Davey JW, Jiggins CD (2015b) Evaluating the use of ABBA-BABA statistics to locate introgressed loci. *Molecular Biology and Evolution*, **32**, 244–257.

Martinsen GD, Whitham TG, Turek RJ, Keim P (2001) Hybrid populations selectively filter gene introgression between species. *Evolution*, **55**, 1325–1335.

Mavárez J, Linares M (2008) Homoploid hybrid speciation in animals. *Molecular Ecology*, **17**, 4181–4185.

Mavárez J, Salazar CA, Bermingham E *et al.* (2006) Speciation by hybridization in *Heliconius* butterflies. *Nature*, **441**, 868–871.

Mayr E (1963) *Animal Species and Evolution*. Harvard University Press, Cambridge, MA.

Mazzuchelli J, Kocher T, Yang F, Martins C (2012) Integrating cytogenetics and genomics in comparative evolutionary studies of cichlid fish. *BMC Genomics*, **13**, 463.

McGaugh SE, Heil CSS, Manzano-Winkler B *et al.* (2012) Recombination modulates how selection affects linked sites in *Drosophila*. *PLoS Biology*, **10**, e1001422.

McMahan CD, Chakrabarty P, Sparks JS, Smith WM, Davis MP (2013) Temporal patterns of diversification across global

cichlid biodiversity (Acanthomorpha: Cichlidae). *PLoS ONE*, **8**, e71162.

Meyer A, Salzburger W, Schartl M (2006) Hybrid origin of a swordtail species (Teleostei: *Xiphophorus clemenciae*) driven by sexual selection. *Molecular Ecology*, **15**, 721–730.

Meyer BS, Matschiner M, Salzburger W (2015) A tribal level phylogeny of Lake Tanganyika cichlid fishes based on a genomic multi-marker approach. *Molecular Phylogenetics and Evolution*, **83**, 56–71.

Meyer BS, Matschiner M, Salzburger W (2016) Disentangling incomplete lineage sorting and introgression to refine species-tree estimates for Lake Tanganyika cichlid fishes. *Systematic Biology*. doi: 10.1093/sysbio/syw069.

Miller JR, Delcher AL, Koren S *et al.* (2008) Aggressive assembly of pyrosequencing reads with mates. *Bioinformatics*, **24**, 2818–2824.

Muschick M, Indermaur A, Salzburger W (2012) Convergent evolution within an adaptive radiation of cichlid fishes. *Current Biology*, **22**, 2362–2368.

Nachman MW, Payseur BA (2012) Recombination rate variation and speciation: theoretical predictions and empirical results from rabbits and mice. *Philosophical Transactions of the Royal Society B: Biological Sciences*, **367**, 409–421.

Nadeau NJ, Martin SH, Kozak KM *et al.* (2013) Genome-wide patterns of divergence and gene flow across a butterfly radiation. *Molecular Ecology*, **22**, 814–826.

Nevado B, Koblmüller S, Sturmbauer C *et al.* (2009) Complete mitochondrial DNA replacement in a Lake Tanganyika cichlid fish. *Molecular Ecology*, **18**, 4240–4255.

Niehuis O, Gibson JD, Rosenberg MS *et al.* (2010) Recombination and its impact on the genome of the haplodiploid parasitoid wasp *Nasonia*. *PLoS ONE*, **5**, e8597.

Nolte AW, Tautz D (2010) Understanding the onset of hybrid speciation. *Trends in Genetics*, **26**, 54–58.

Norris LC, Main BJ, Lee Y *et al.* (2015) Adaptive introgression in an African malaria mosquito coincident with the increased usage of insecticide-treated bed nets. *Proceedings of the National Academy of Sciences*, **112**, 815–820.

Pardo-Diaz C, Salazar C, Baxter SW *et al.* (2012) Adaptive introgression across species boundaries in *Heliconius* butterflies. *PLoS Genetics*, **8**, e1002752.

Parra G, Bradnam K, Korf I (2007) CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics*, **23**, 1061–1067.

Parra G, Bradnam K, Ning Z, Keane T, Korf I (2009) Assessing the gene space in draft genomes. *Nucleic Acids Research*, **37**, 289–297.

Patten MM, Carioscia SA, Linnen CR (2015) Biased introgression of mitochondrial and nuclear genes: a comparison of diploid and haplodiploid systems. *Molecular Ecology*, **24**, 5200–5210.

Paun O, Forest F, Fay MF, Chase MW (2009) Hybrid speciation in angiosperms: parental divergence drives ploidy. *New Phytologist*, **182**, 507–518.

Pease JB, Hahn MW (2013) More accurate phylogenies inferred from low-recombination regions in the presence of incomplete lineage sorting. *Evolution*, **67**, 2376–2384.

Phillips D, Jenkins G, Macaulay M *et al.* (2015) The effect of temperature on the male and female recombination landscape of barley. *New Phytologist*, **208**, 421–429.

Poelstra JW, Vijay N, Bossu CM *et al.* (2014) The genomic landscape underlying phenotypic integrity in the face of gene flow in crows. *Science*, **344**, 1410–1414.

Pollard DA, Iyer VN, Moses AM, Eisen MB (2006) Widespread discordance of gene trees with species tree in drosophila: evidence for incomplete lineage sorting. *PLoS Genetics*, **2**, 1634–1647.

Rambaut A, Suchard MA, Xie D, Drummond AJ (2014) Tracer v1.6.

Rieseberg LH, Whitton J, Gardner K (1999) Hybrid zones and the genetic architecture of a barrier to gene flow between two sunflower species. *Genetics*, **152**, 713–727.

Rieseberg LH, Raymond O, Rosenthal DM *et al.* (2003) Major ecological transitions in wild sunflowers facilitated by hybridization. *Science*, **301**, 1211–1216.

Rockman MV, Kruglyak L (2009) Recombinational landscape and population genomics of Caenorhabditis elegans. *PLoS Genetics*, **5**, e1000419.

Roesti M, Hendry AP, Salzburger W, Berner D (2012) Genome divergence during evolutionary diversification as revealed in replicate lake-stream stickleback population pairs. *Molecular Ecology*, **21**, 2852–2862.

Roesti M, Moser D, Berner D (2013) Recombination in the threespine stickleback genome-patterns and consequences. *Molecular Ecology*, **22**, 3014–3027.

Roesti M, Gavrilets S, Hendry AP, Salzburger W, Berner D (2014) The genomic signature of parallel adaptation from shared genetic variation. *Molecular Ecology*, **23**, 3944–3956.

Roesti M, Kueng B, Moser D, Berner D (2015) The genomics of ecological vicariance in threespine stickleback fish. *Nature Communications*, **6**, 8767.

Rosenzweig BK, Pease JB, Besansky NJ, Hahn MW (2016) Powerful methods for detecting introgressed regions from population genomic data. *Molecular Ecology*, **25**, 2387–2397.

Rüber L, Meyer A, Sturmbauer C, Verheyen E (2001) Population structure in two sympatric species of the Lake Tanganyika cichlid tribe Eretmodini: evidence for introgression. *Molecular Ecology*, **10**, 1207–1225.

Sætre G-P (2013) Hybridization is important in evolution, but is speciation? *Journal of Evolutionary Biology*, **26**, 256–258.

Salazar C, Baxter SW, Pardo-Diaz C *et al.* (2010) Genetic evidence for hybrid trait speciation in *Heliconius* butterflies. *PLoS Genetics*, **6**, e1000930.

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

Salzburger W, Baric S, Sturmbauer C (2002) Speciation via introgressive hybridization in East African cichlids? *Molecular Ecology*, **11**, 619–625.

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 Evolutionary Biology*, **5**, 17.

Salzburger W, Van BB, Cohen AS (2014) Ecology and evolution of the African Great Lakes and their faunas. *Annual Review of Ecology, Evolution, and Systematics*, **45**, 519–545.

Santos ME, Salzburger W (2012) Evolution. How cichlids diversify. *Science (New York, New York)*, **338**, 619–621.

Schelly R, Salzburger W, Koblmüller S, Duftner N, Sturmbauer C (2006) Phylogenetic relationships of the lamprologine cichlid genus *Lepidiolamprologus* (Teleostei: Perciformes) based on

mitochondrial and nuclear sequences, suggesting introgressive hybridization. *Molecular Phylogenetics and Evolution*, **38**, 426–438.

Scholz CA, Johnson TC, Cohen AS *et al.* (2007) East African megadroughts between 135 and 75 thousand years ago and bearing on early-moderns human origins. *Proceedings of the National Academy of Sciences*, **104**, 16416–16421.

Schumer M, Cui R, Boussau B *et al.* (2013) An evaluation of the hybrid speciation hypothesis for *Xiphophorus clemenciae* based on whole genome sequences. *Evolution*, **67**, 1155–1168.

Schwarzer J, Swartz ER, Vreven E *et al.* (2012) Repeated transwatershed hybridization among haplochromine cichlids (Cichlidae) was triggered by Neogene landscape evolution. *Proceedings of the Royal Society B: Biological Sciences*, **279**, 4389–4398.

Seehausen O (2004) Hybridization and adaptive radiation. *Trends in Ecology and Evolution*, **19**, 198–207.

Seehausen O (2006) African cichlid fish: a model system in adaptive radiation research. *Proceedings. Biological Sciences/The Royal Society*, **273**, 1987–1998.

Sella G, Petrov DA, Przeworski M, Andolfatto P (2009) Pervasive natural selection in the *Drosophila* genome? *PLoS Genetics*, **5**, e1000495.

Smadja CM, Butlin RK (2011) A framework for comparing processes of speciation in the presence of gene flow. *Molecular Ecology*, **20**, 5123–5140.

Sousa-Santos C, Gante HF, Robalo J *et al.* (2014) Evolutionary history and population genetics of a cyprinid fish (*Iberochondrostoma olisiponensis*) endangered by introgression from a more abundant relative. *Conservation Genetics*, **15**, 665–677.

Stamatakis A (2014) RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, **30**, 1312–1313.

Stiassny MLJ (1997) A phylogenetic overview of the lamprologine cichlids of Africa (Teleostei, Cichlidae): a morphological perspective. *South African Journal of Science*, **93**, 513–523.

Sturmbauer C, Baric S, Salzburger W, Rüber L, Verheyen E (2001) Lake level fluctuations synchronize genetic divergences of cichlid fishes in African lakes. *Molecular Biology and Evolution*, **18**, 144–154.

Sturmbauer C, Salzburger W, Duftner N, Schelly R, Koblmüller S (2010) Evolutionary history of the Lake Tanganyika cichlid tribe Lamprologini (Teleostei: Perciformes) derived from mitochondrial and nuclear DNA data. *Molecular Phylogenetics and Evolution*, **57**, 266–284.

Taborsky M (1984) Broodcare helpers in the cichlid fish *Lamprologus brichardi*: their costs and benefits. *Animal Behaviour*, **32**, 1236–1252.

Taborsky M, Limberger D (1981) Helpers in fish. *Behavioral Ecology and Sociobiology*, **8**, 143–145.

Tavaré S (1986) Some probabilistic and statistical problems in the analysis of DNA sequences. *Lectures on Mathematics in the Life Sciences*, **17**, 57–86.

Than C, Ruths D, Nakhleh L (2008) PhyloNet: a software package for analyzing and reconstructing reticulate evolutionary relationships. *BMC Bioinformatics*, **9**, 322.

Tigano A, Friesen VL (2016) Genomics of local adaptation with gene flow. *Molecular Ecology*, **25**, 2144–2164.

Tine M, Kuhl H, Gagnaire P-A *et al.* (2014) European sea bass genome and its variation provide insights into adaptation to euryhalinity and speciation. *Nature Communications*, **5**, 5770.

Tortereau F, Servin B, Frantz L *et al.* (2012) A high density recombination map of the pig reveals a correlation between sex-specific recombination and GC content. *BMC Genomics*, **13**, 586.

Trier CN, Hermansen JS, Sætre G-P, Bailey RI (2014) Evidence for mito-nuclear and sex-linked reproductive barriers between the hybrid Italian sparrow and its parent species. *PLoS Genetics*, **10**, e1004075.

Turner GF, Seehausen O, Knight ME, Allender CJ, Robinson RL (2001) How many species of cichlid fishes are there in African lakes? *Molecular Ecology*, **10**, 793–806.

Turner TL, Hahn MW, Nuzhdin SV (2005) Genomic islands of speciation in *Anopheles gambiae*. *PLoS Biology*, **3**, e285.

Wagner CE, Harmon LJ, Seehausen O (2012) Ecological opportunity and sexual selection together predict adaptive radiation. *Nature*, **487**, 366–369.

Wagner CE, Keller I, Wittwer S *et al.* (2013) Genome-wide RAD sequence data provide unprecedented resolution of species boundaries and relationships in the Lake Victoria cichlid adaptive radiation. *Molecular Ecology*, **22**, 787–798.

Webster MT, Hurst LD (2012) Direct and indirect consequences of meiotic recombination: implications for genome evolution. *Trends in Genetics*, **28**, 101–109.

Wong M, Balshine S (2011) The evolution of cooperative breeding in the African cichlid fish, *Neolamprologus pulcher*. *Biological Reviews of the Cambridge Philosophical Society*, **86**, 511–530.

Wu C-I (2001) The genic view of the process of speciation. *Journal of Evolutionary Biology*, **14**, 851–865.

Yatabe Y, Kane NC, Scotti-Saintagne C, Rieseberg LH (2007) Rampant gene exchange across a strong reproductive barrier between the annual sunflowers. *Helianthus annuus and H. petiolaris*. *Genetics*, **175**, 1883–1893.

Yu Y, Dong J, Liu KJ, Nakhleh L (2014) Maximum likelihood inference of reticulate evolutionary histories. *Proceedings of the National Academy of Sciences*, **111**, 16448–16453.

Zamani N, Russell P, Lantz H *et al.* (2013) Unsupervised genome-wide recognition of local relationship patterns. *BMC Genomics*, **14**, 347.

H.F.G., S.J. and W.S. designed the research; H.F.G. and Ma.M. performed wet laboratory analysis; Ma.M. involved in genome assembly and whole-genome mapping; Mi.M. and H.F.G. involved in statistical analyses; H.F.G., Mi.M., and Ma.M. wrote manuscript with input from all coauthors.

Data accessibility

Raw sequencing reads associated with this study are available at the European Nucleotide Archive's Short Read Archive (http://www.ebi.ac.uk/ena; primary and secondary accession nos PRJEB12322 and ERP013786). RAxML trees and alignment files are available at Dryad Repository (doi:10.5061/dryad.jr67t). An 'Introgression tutorial' focusing on the implementation of

phylogenomic methods (SAGUARO, BEAST and PHYLONET) by Mi.M is available at https://github.com/mmatschiner/Introgression-Tutorial.

Supporting information

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

Fig. S1 Distribution of mapping coverage across the five *Neolamprologus* and *M. zebra*.

Fig. S2 Frequencies of local topologies in three-taxon comparisons.

Fig. S3 Distribution of chromosomal segments supporting the three most frequent unrooted local topologies.

Fig. S4 Bias in the genomic distribution of introgression and incomplete lineage sorting.

Table S1 Locality information of sequenced species.

Table S2 Coverage and numbers of partial and complete core genes for each of the newly sequenced genomes.

Table S3 Number of fixed and polymorphic single nucleotide polymorphisms across the five *Neolamprologus* and *M. zebra* relative to *O. niloticus*.

Table S4 Numbers of chromosomal segments, mean Bayesian posterior probabilities (BPP) and median ages of alternative rooted topologies of five taxa identified by BEAST.

Table S5 Frequencies of local topologies in three-taxon comparisons and hypotheses of incomplete lineage sorting and introgression.

Table S6 Numbers of chromosomal segments and sites, and median length of segments supporting alternative unrooted topologies of five taxa identified by SAGUARO.

Table S7 ABBA-BABA genome-wide tests of shared variants in four-taxon comparisons using Patterson's *D* statistic.

Table S8 Maximum-likelihood tests of reticulation using PHYLONET.

Table S9 Ages (in million years) of all clades and inferred introgression events.