Received Date : 15-Mar-2016
Revised Date : 15-Aug-2016
Accepted Date : 22-Aug-2016
Article type : Special Issue

Demographic modeling with whole genome data reveals parallel origin of similar *Pundamilia* cichlid species after hybridization

Joana I.MEIER (JIM)[1,2,3,]*, Vitor C. SOUSA (VCS)[2,4], David A. MARQUES (DAM)[1,2,3], Oliver M. SELZ (OMS)[1,3], Catherine E. WAGNER (CEW)[5], Laurent EXCOFFIER (LE)[2,4] & Ole SEEHAUSEN (OS)[1,3]

[1]Aquatic Ecology and Evolution, Institute of Ecology and Evolution, University of Bern, Baltzerstrasse 6, CH-3012 Bern, Switzerland; [2]CMPG, Institute of Ecology and Evolution, University of Bern, Baltzerstrasse 6, CH-3012 Bern, Switzerland; [3]Department of Fish Ecology and Evolution, EAWAG Swiss Federal Institute of Aquatic Science and Technology, Center for Ecology, Evolution and Biogeochemistry, Seestrasse 79, CH-6047 Kastanienbaum, Switzerland; [4]Swiss Institute of Bioinformatics, CH-1015 Lausanne, Switzerland; [5]Biodiversity Institute & Department of Botany, University of Wyoming, Laramie, WY USA

Keywords: ecological speciation, cichlid fish, adaptive radiation, demographic modeling, hybrid speciation

*Corresponding author: Joana Isabel Meier, Baltzerstrasse 6, CH-3012 Bern, Switzerland, +41 31 631 3016, joana.meier@iee.unibe.ch

Running title: Hybrid parallel origin of cichlid species

Abstract

Modes and mechanisms of speciation are best studied in young species pairs. In older taxa it is increasingly difficult to distinguish what happened during speciation from what happened after speciation. Lake Victoria cichlids in the genus *Pundamilia* encompass a complex of young species and polymorphic populations. One *Pundamilia* species pair, *P. pundamilia* and *P. nyererei*, is particularly well-suited to study speciation because sympatric population pairs occur with different levels of

phenotypic differentiation and reproductive isolation at different rocky islands within the lake. Genetic distances between allopatric island populations of the same nominal species often exceed those between the sympatric species. It thus remained unresolved whether speciation into *P. nyererei* and *P. pundamilia* occurred once, followed by geographical range expansion and interspecific gene flow in local sympatry, or if the species pair arose repeatedly by parallel speciation. Here we use genomic data and demographic modeling to test these alternative evolutionary scenarios. We demonstrate that gene flow plays a strong role in shaping the observed patterns of genetic similarity, including both gene flow between sympatric species and gene flow between allopatric populations, as well as recent and early gene flow. The best supported model for the origin of *P. pundamilia* and *P. nyererei* population pairs at two different islands is one where speciation happened twice, whereby the second speciation event follows shortly after introgression from an allopatric *P. nyererei* population that arose earlier. Our findings support the hypothesis that very similar species may arise repeatedly, potentially facilitated by introgressed genetic variation.

Introduction

Young species pairs are especially valuable in speciation research because differences between them can be attributed with some confidence to processes operating during speciation. In older species with complete reproductive isolation it is difficult to distinguish factors that contributed to reproductive isolation and divergence during speciation from those that only occurred after speciation (Coyne and Orr 2004). However, most young sister species pairs have emerged from singular evolutionary events and hence provide only one snapshot in the presumed continuum between a single panmictic population and two reproductively isolated species (Nosil and Feder 2013; Faria, et al. 2014; Seehausen, et al. 2014).

Natural replicates, such as similar species pairs that have evolved under similar contrasting selection regimes (Schluter and Nagel 1995) allow for powerful tests of the underlying processes leading to speciation, and the role of natural and sexual selection in speciation (Schluter and Nagel 1995; Jones, et al. 2012; Nosil 2012; Butlin, et al. 2014). Consistent phenotype-environment correlations suggest the repeated action of ecology in shaping trait values, and traits or loci that differ consistently between species in replicated species pairs are strong candidates for divergent selection and/or involvement in reproductive isolation (Schluter and Nagel 1995; Rundle, et al. 2000; Schluter, et al. 2004; Faria, et al. 2014). In many studies of repeated evolution, the genetic basis of traits underlying similar phenotypes is unknown. Nevertheless, the traits that evolved repeatedly in parallel are often assumed to have arisen independently through separate *de novo* mutations (narrow sense definition of parallel evolution) but such mutations could also have been recruited from shared ancestral polymorphisms or interspecific gene flow (Schluter and Conte 2009; Johannesson, et al. 2010; Faria, et al. 2014). The distinction between "parallel evolution" and "convergent evolution" (independent evolution of similar phenotypes in distantly related lineages or based on different pathways) is disputed (Haas and Simpson 1946; Arendt and Reznick 2008; Conte, et al. 2012). Here, we use the term parallel evolution as our study species are very closely related. A special form of parallel evolution, parallel speciation, is when reproductive isolation (RI) in several speciation events also evolves in parallel, driven by the same mechanism. In this case, species are reproductively isolated from their sister species, but would potentially lack RI against independently evolved, geographically distant species experiencing similar selection environments and expressing similar phenotypes (Schluter and Nagel 1995). Parallel evolution of reproductive isolation based on the same traits may occur through *de novo* mutations, parallel sorting of standing genetic variation, or introgressed alleles. In all of these cases, there are independent replicates of the speciation event, even though the alleles involved may have arisen only once and may be shared between the speciation events.

Another type of evolutionary replication, applying not to the origin of species but to the maintenance of species differences, is when a young species pair hybridizes at several independently established secondary contact zones (or geographically isolated instances of breakdown of reproductive isolation between species in primary contact e.g. due to habitat disturbance), and shows parallel maintenance of species differences despite gene flow (Barton and Hewitt 1985; Rand and Harrison 1989; Abbott, et al. 2013). Genetic regions that show reduced introgression in independent hybrid zones likely contain loci that are under divergent selection, or contribute to reproductive isolation between the species (Barton and Hewitt 1989; Rieseberg, et al. 1999; Rieseberg and Buerkle 2002; Gompert and Buerkle 2009; Teeter, et al. 2010; Harrison and Larson 2016). Studying such situations allows investigation of the genomic architecture of adaptation and reproductive isolation, the types of incompatibilities and the roles of different types of selection. If a species pair is young, then the mechanisms that can be inferred are likely to have played a role in initiating speciation too, whereas this inference cannot be made when hybrid zones between older species are being studied (Nosil and Schluter 2011).

Systems with either parallel speciation or parallel maintenance of species differences in multiple hybrid zones are thus both useful for studying the processes underlying speciation. However, distinguishing between these scenarios is not trivial as they can both produce stronger neutral genetic clustering by geographical proximity than by phenotypic similarity (Johannesson, et al. 2010; Bierne, et al. 2013; Faria, et al. 2014). It is crucial to know the evolutionary history of these natural replicates to assess the level of evolutionary independence and to study the underlying evolutionary mechanisms, including natural and sexual selection, *de novo* mutations, gene flow, drift, and standing genetic variation. Importantly, parallel speciation has been taken to imply that speciation is to some extent a predictable outcome of adaptation, whereas parallel maintenance of species differences does not allow this inference (Stern and Orgogozo 2009; Rosenblum, et al. 2014). Genetic clustering of species and ecotypes by geographic proximity rather than by phenotypic similarity has been shown in species complexes of sunflowers (Renaut, et al. 2014), sticklebacks

(Rundle, et al. 2000), *Littorina* periwinkle snails (Butlin, et al. 2014; Ravinet, et al. 2015), Darwin finches (Lamichhaney, et al. 2015), *Timema* stick insects (Nosil, et al. 2002), and cichlid fish of Lake Malawi (Allender, et al. 2003) and Victoria (Seehausen, et al. 2008; Konijnendijk, et al. 2011; Magalhaes, et al. 2012). However, tests discriminating single and multiple origins of ecotypes or species in the face of continuing gene flow are often lacking (but see Butlin, et al. 2014). Here, we explicitly test alternative scenarios for speciation in *Pundamilia* cichlids using genomic data and demographic modeling.

Lake Victoria haplochromine cichlids are an adaptive radiation of approximately 500 endemic species that evolved over the past 15,000 years (Seehausen 2006). They comprise multiple replicates of young species pairs differing along the same dominant axes of phenotypic and ecological divergence, most strikingly in male nuptial coloration and water depth occupation, whereas less closely related species can differ dramatically also in morphology and ecology (Seehausen 2015). The Lake Victoria endemic genus *Pundamilia* offers a unique opportunity to study speciation in natural replicates with varying extents of reproductive isolation, of what appears to be the same sympatric species pair at different sites within the lake. The sister species *Pundamilia pundamilia* and *P. nyererei* coexist in complete sympatry at many sites in Lake Victoria (Fig. 1), including islands in the Mwanza Gulf, an environment with considerable variation in water turbidity between sites. At islands with clear water, *P. pundamilia* and *P. nyererei* are distinct species with highly divergent male nuptial coloration, some differences in ecology and morphology, and strong assortative mating (Fig. 1, Table S1, Seehausen 2009).

Pundamilia cichlids became a model system for studying the interaction of sexual and natural selection in speciation and the role of sensory drive mechanisms, as visual system adaptation to contrasting light environments between the sympatric sister species is associated with divergent sexual selection on male nuptial coloration (Maan, et al. 2006; Seehausen, et al. 2008). In Lake Victoria, particulate matter selectively absorbs and scatters short-wavelength light (Okullo, et al.

2007) shifting the light gradient from broad spectrum daylight in shallow water towards more red light in deeper water. This light gradient is steeper in more turbid water and less steep in clear waters (Seehausen, et al. 1997). *P. nyererei,* which lives in deeper water with a relatively more red-shifted light spectrum, has bright red male nuptial coloration and visual sensitivity shifted towards red light relative to *P. pundamilia* (Table S1, Maan, et al. 2006). In addition, *P. nyererei* females, cryptically yellow-brown themselves, choose their mates based on nuptial coloration, preferring conspecific red over heterospecific blue males (Seehausen, et al. 1997; Seehausen and van Alphen 1998; Selz, Pierotti, et al. 2014) and have a preference for redder males (Maan, et al. 2004; Stelkens, et al. 2008; Maan, et al. 2010; Selz, et al. 2016). In contrast, the more shallow breeding *P. pundamilia* has blue male nuptial coloration (Seehausen 1996), visual sensitivity that is more shifted towards blue light (Maan, et al. 2006), and females prefer conspecific blue over heterospecific red males (Table S1, Seehausen, et al. 1997; Seehausen and van Alphen 1998; Selz, Pierotti, et al. 2014). The two species also differ in many other traits such as female coloration, and multiple morphological and ecological characteristics (Table S1).

In this study, we focus on species pairs from islands with different water turbidity and hence different steepness of the ambient light gradient. Higher turbidity and steeper light gradients are associated with weaker genetic differentiation between *P. pundamilia* and *P. nyererei* (Seehausen, et al. 2008). Whereas the species are clearly distinct at the clear water Makobe Island in the open lake just north of the Mwanza Gulf, hybrids can be seen regularly at the intermediately turbid sites of Kissenda Island and Python Island in the Mwanza Gulf (Fig. 1). Further south in the Gulf, at Luanso Island, the water is even more turbid and *Pundamilia* males vary extensively in nuptial coloration, including clearly red and clearly blue individuals, but intermediate phenotypes are most abundant and genetic data is consistent with a single panmictic population (Seehausen, et al. 2008; van der Sluijs, et al. 2008). Likewise, female preference for male coloration has a bimodal frequency

distribution at the islands where male coloration is bimodally distributed, but has a unimodal frequency distribution at Luanso Island where most females show no behavioral mating preference and few females show preferences for either red or blue males (Seehausen, et al. 2008; van der Sluijs, et al. 2008).

Populations of *P. nyererei* exhibit considerable phenotypic variation both in morphology and in chroma (*i.e.* purity and intensity of colour) that can be classified into two types (Seehausen 1996). One type occurs mainly in the western Mwanza Gulf (e.g. Python and Kissenda Island), whereas the other type occurs in the eastern Mwanza Gulf and in the main body of Lake Victoria (e.g. Makobe Island, Fig. 1, Seehausen 1996). They differ mainly in body depth, jaw length, and female and male coloration (Fig. 1, Table S1, Seehausen 1996). *P. nyererei* occupies the entire depth range at Kissenda and Python Islands, whereas at Makobe Island it lives almost exclusively deeper than 3 m (Seehausen 1997; Seehausen, et al. 2008). The *P. pundamilia* populations used in our study are similar in coloration (Castillo Cajas, et al. 2012) but they differ in relative head length, eye size and in the number and shape of the vertical bars (Table S1, Seehausen 1996; Seehausen, et al. 1998). These population differences may have arisen through independent evolution after speciation and range expansion, or they may suggest that populations allocated to the same nominal species evolved more than once in parallel.

Seehausen, et al. (2008) found that neutral genetic divergence (based on microsatellite data) between sympatric *P. pundamilia* and *P. nyererei* within an island was often lower than that between allopatric populations belonging to the same morphologically defined nominal species. This observation raised the question of whether interspecific gene flow in sympatry had eroded the phylogenetic signature of common ancestry of all *P. nyererei*, or if similar red and blue *Pundamilia* species pairs had evolved in parallel at different islands. Whereas *P. pundamilia* (blue) is found at almost every sampled rocky habitat patch in Lake Victoria (Seehausen, et al. 1998) and has the highest record of sympatry with other *Pundamilia* species (Seehausen 1996; Seehausen and van

Alphen 1999), the distribution range of *P. nyererei* (red) is patchy, and *P. nyererei* populations are only found at sites where *P. pundamilia* is also present (Seehausen and van Alphen 1999). It has thus been proposed that *P. pundamilia*-type fish (blue males) may represent the ancestral form that colonized the different islands and rocky mainland shores of Lake Victoria (Seehausen 1996, 1997; Seehausen and Schluter 2004; Dijkstra, et al. 2007). Populations of *P. nyererei* (red males) occur only at some islands, with a patchy geographical distribution, and may have evolved independently at different islands or, alternatively, *P. nyererei* may have split once from *P. pundamilia* and subsequently colonized the different islands where it experienced gene flow with the resident *P. pundamilia* populations (Seehausen 1996, 1997; Seehausen and Schluter 2004; Dijkstra, et al. 2007).

Distinguishing these evolutionary scenarios has important consequences for understanding the speciation process and the effects of environmental variation on speciation and persistence of sister species. Here we test alternative evolutionary scenarios including, single, parallel, and hybrid origin models using whole genome and partial genome sequences of population samples, and demographic modeling.

Materials and Methods

Samples

Wild males of *P. nyererei* and *P. pundamilia* were collected with gill nets and by angling at Makobe Island and at three islands in the Mwanza Gulf (Kissenda, Python and Luanso Islands, Fig. 1) in 2005 and in 2010. Phenotypically intermediate males were also sampled at islands where they occurred (see Table S2 for more sample information). DNA was extracted from fin clips using a standard phenol-chloroform protocol (Sambrook and Russell 2001).

RAD sequencing

Restriction-site Associated DNA sequencing (RADseq) was performed following a standard protocol (Baird, et al. 2008) with minor modifications. Restriction digestion was done overnight using the restriction endonuclease HF-*Sbf*I (NewEngland Biolabs) and 1 µg DNA per sample. P1 adapters contained 5-8 bp long barcodes differing by at least two nucleotides from all other barcodes. The DNA was sheared with a Covaris S220 Focused-Ultra sonicator and fragments of 300 – 600 bp length were manually cut from an agarose gel. All libraries were single-end sequenced on an Illumina HiSeq 2500 sequencer. The reads were demultiplexed and trimmed to 84 bp with the process_radtags script from the Stacks pipeline (Catchen, et al. 2013), correcting single errors in the barcode and the restriction site, and discarding reads with incomplete restriction sites. The FastX toolkit (http://hannonlab.cshl.edu/fastx_toolkit) was used to remove all reads containing at least one base with a Phred quality score below 10 and reads with more than 10% of bases with quality less than 30. The reads of each individual were then mapped to the *Pundamilia nyererei* reference genome (Brawand, et al. 2014) using Bowtie2 (Langmead and Salzberg 2012) with the end-to-end alignment option and default parameters. Base score recalibration was performed using empirical error rate estimations derived from bacteriophage PhiX reads sequenced together with the RAD libraries following Marques, et al. (2016). Single nucleotide polymorphisms (SNPs) and genotypes were called using GATK Unified Genotyper v. 3.5 (McKenna, et al. 2010). All sites were then filtered with a custom-made Python script, bcftools v. 0.1.12, and vcftools v. 0.1.14 (Danecek, et al. 2011). We removed sites within 5 bp from indels (insertions/deletions) to avoid false SNPs due to potential misalignment problems. Genotypes were required to have a depth of coverage of at least ten reads and a quality value (GQ) of 30. Sites with more than 25% missing data were removed. Sites with a mean genotype depth greater than 59.6x (1.5 times the interquartile range from the mean) were excluded as they are expected to be enriched for paralogs. Indeed, 25% of the SNPs with minor allele frequency of at least 0.05 among these sites were heterozygous in over 80% of the individuals.

Genetic distances

To study population structure among our populations we performed Principal Components Analysis (PCA) with the R-package SNPRelate (Zheng, et al. 2012) using biallelic SNPs from the RAD sequencing data set with a minor allele frequency of at least 5% over all sequenced individuals. A dataset of unlinked SNPs was obtained by pruning SNPs that were in high linkage disequilibrium (LD) across all populations combined. We used plink ('-indep-pairwise 50 5 0.5', Purcell, et al. 2007) to calculate LD (r^2) between each pair of SNPs in a window of 50 SNPs. If a pair of SNPs had r^2 greater than 0.5, one of the two SNPs was removed at random. The window was then shifted by 5 SNPs and the procedure was repeated until the last SNP was reached.

To measure genetic distances between the populations, we created a RAD dataset without missing data by randomly sampling 12 individuals of each population per site. Individuals phenotypically identified as putative hybrids were excluded. We used all biallelic sites in which the rarer allele was observed at least three times across all individuals (minor allele frequency, maf 1.5%) to calculate pairwise F_{ST} values. We applied a Mantel test with 1,000 permutations to assess if genetic distances are correlated with waterway distances between the islands using Arlequin v. 3.5.2.3 (Excoffier and Lischer 2010).

Whole-genome sequencing

Whole-genome sequencing data was generated for a subset of the same individuals (four individuals of *P. pundamilia* and four of *P. nyererei* each from Makobe and Python Islands, see Table S2) using PCR-free library preparation (Kozarewa, et al. 2009) and Illumina HiSeq 3000 paired-end sequencing. To avoid any sequencing lane effects and to get an even read representation, all individuals were given separate barcodes and subsequently sequenced together on four Illumina HiSeq 3000 lanes. Local alignment against the *Pundamilia nyererei* reference genome (Brawand, et al. 2014) was

performed with Bowtie 2 (Langmead and Salzberg 2012), and variant calling and genotyping with Haplotype Caller (GATK v. 3. 5, McKenna, et al. 2010). To avoid potential paralogous regions, we excluded sites with a mean depth per individual greater than 1.5 times the interquartile range from the mean and we excluded sites deviating from Hardy Weinberg Equilibrium (HWE) due to excess observed heterozygosity over all 16 individuals pooled using vcftools v. 0.1.14 (Danecek, et al. 2011) and a p-value cutoff of 0.001. We did not have enough power to filter for deviation from HWE in single populations. All sites were required to have a depth of coverage of at least 15 reads in each individual. All biallelic SNPs were then extracted for demographic modeling.

In order to assess if the individuals used for whole-genome sequencing differed in ancestry proportions from the other individuals of their respective populations in the RAD dataset, we performed Bayesian clustering assignment of all individuals with two to seven clusters (K=2-7) using STRUCTURE v2.34 (Pritchard, et al. 2000). We used all bi-allelic sites with a minor allele frequency of 5% in the RAD dataset. We ran 10 replicates each assuming two to seven clusters with 500,000 burnin and 1,000,000 sampling steps. The most likely number of clusters was identified by the highest delta K among all runs (Evanno, et al. 2005) with Structure Harvester (Earl and vonHoldt 2012). We used the *greedy* algorithm within clumpp v1.1.2 (Jakobsson and Rosenberg 2007) to account for variation in cluster labels across the 10 replicates for each number of clusters.

Demographic modeling

In order to discriminate among alternative evolutionary scenarios for the origin of the *P. nyererei* and *P. pundamilia* populations and to estimate relevant parameters, we used demographic modeling based on the site frequency spectrum (SFS) of the whole-genome sequencing data. First, we used only the populations at Makobe Island, where the two species are most distinct, to assess if speciation occurred in a period of geographical isolation or in the face of gene flow (Fig. 3). Next, we

contrasted alternative demographic models for the evolution of the species pairs at Makobe (representing islands in the open lake) and Python (representing islands in the western Mwanza Gulf) combined. We compared models of single, parallel, and hybrid origin of the red and the blue species, each combined with different post-speciation gene flow scenarios (Fig. 4).

For each model, the fit to the observed multidimensional SFS was maximized using the composite-likelihood method implemented in fastsimcoal v. 2.5.2 (Excoffier, et al. 2013) with the following options: -N 100,000 (number of coalescent simulations), -L 40 (number of expectation-maximization (EM) cycles), -M 0.001 (minimum relative difference in parameter values for the stopping criterion), and -C 10 (minimum observed SFS entry count taken into account for parameter estimation). For all model parameters we used wide search ranges with log-uniform distributions (see Table S3 and S4). For each demographic model we performed 500 independent fastsimcoal runs to determine the parameter estimates leading to the maximum likelihood (Excoffier, et al. 2013).

The observed multi-dimensional SFS was computed with Arlequin v. 3.5.2.3 (Excoffier and Lischer 2010). As we currently lack a good estimate for the mutation rate in Lake Victoria cichlids, all parameter estimates are relative to the time of the first split in each model, which was fixed to be 6,000 generations before present. Although the real splitting time between *P. pundamilia* and *P. nyererei* is likely more recent, this initial splitting time provides an upper bound, as it roughly coincides with the refilling of the Lake Victoria basin after several thousand years of complete desiccation (Johnson, et al. 2000), assuming a generation time of two years. Note that the species pair and the entire genus *Pundamilia* are endemic to Lake Victoria itself and are absent from the other smaller lakes in the region. The species can only have arisen after the lake filled again with water. Many polymorphisms will not have arisen in the past 6,000 generations but rather represent ancestral standing variation. In the demographic models the levels of ancestral standing variation are represented by the effective size of the ancestral population.

The best fitting demographic model was identified on the basis of the Akaike's information criterion (AIC, Akaike 1974). However, we note that given the presence of linked sites in our data, the reported AIC values should be interpreted with caution (Excoffier, et al. 2013). For this reason, we also examined the likelihood distributions obtained based on 100 expected SFS, each approximated using 1 million coalescent simulations under the parameters that maximize the likelihood for each model. These distributions inform us about the variance in likelihoods estimated by fastsimcoal2. An overlap of these distributions between models would indicate no significant difference between the fit of alternative models as the difference may be attributed to the variance in the SFS approximation.

To get confidence intervals for the parameter estimates, we used a non-parametric block-bootstrap approach. To account for linkage disequilibrium, 200 bootstrap datasets were obtained by dividing the SNPs into 100 blocks, and sampling with replacement 100 blocks for each bootstrap data set, to match the original dataset size. The parameter point estimates from the run with the highest likelihood (out of 100 independent runs) from each bootstrapping replicate were then used to compute 95 percentile confidence intervals with the R-package "boot" (Canty and Ripley 2015)

Results

Datasets

The total width of the data matrix derived from RAD sequencing, after alignment to the reference genome and genotype calling, was 3,102,689 bp (including invariant sites). After all filtering steps, we retained 21,090 SNPs with maximum 25% missing data at a minor allele frequency level of 0.01. The mean proportion of missing data in the filtered data set was 8.7% in the whole RADseq dataset and the mean depth of coverage was 40x.

The whole-genome sequencing dataset after mapping to the *Pundamilia nyererei* reference genome contained a total of 653 million bp with a minimum depth of coverage of 5x and a maximum of 50% missing data. The mean missing data proportion was 1.5% and average depth of coverage was 19x. The dataset used for the demographic modeling contained 433,893 biallelic SNPs without missing data and a minimum depth of coverage of 15x in all 16 individuals.

Population structure

The RADseq dataset confirms previous finding by Seehausen, et al. (2008) that *P. nyererei* and *P. pundamilia* populations coexisting on the same island are mostly genetically more similar to each other than are allopatric populations of the same nominal species from different islands (Fig. 2, Fig. S1). Interspecific F_{ST} values are highest at the Makobe clear water site, intermediate at Kissenda and Python and zero at Luanso Island, the most turbid site (Fig. 2a, Fig. S1). Intraspecific F_{ST} values among island populations of *P. pundamilia* correlate positively with geographic distance (Mantel test, p-value=0.048), whereas the correlation is non-significant for *P. nyererei* (p-value=0.176), fully consistent with earlier microsatellite-based inference (Seehausen, et al. 2008).

In the PCA analysis, the first component (PC1) carries a signature of geographic distance between islands but also differentiates the sympatric species within islands. PC2 differentiates the sympatric species but the species diverge in opposite directions at Makobe versus at Kissenda and Python Islands. The third principal component groups populations by nominal species (Fig. 2c). Individuals from Luanso Island, thought to be a population of intermediates or natural hybrids, form a single cluster, intermediate between the *P. pundamilia* and *P. nyererei* regions in PCA space, without evidence of color-associated genetic differentiation. On PC1, Luanso individuals form a distinct cluster clearly separated from all other islands. At higher PC axes, all Luanso individuals form a

cluster in the center of the PCA space. Several individuals from Python and Kissenda that were phenotypically identified as putative hybrids are located between the clusters of *P. nyererei* and *P. pundamilia* populations of their respective islands.

Demographic modeling

To test if *P. nyererei* and *P. pundamilia* speciated in allopatry (*i.e*. without gene flow) or in the face of gene flow, and whether gene flow changed over time, we contrasted different scenarios of speciation-with-gene-flow and without gene flow using the Makobe populations because these were least affected by recent changes in water clarity. We compared five models with, 1) secondary contact (only recent gene flow) 2) only early gene flow, 3) different amounts of early and recent gene flow, 4) constant gene flow, and 5) no gene flow (Fig. 3). The time of change in gene flow (*i.e.* start of recent or end of early gene flow) was estimated as a model parameter and allowed to range from 1 to 6,000 generations. Parameter estimates and likelihoods are given in Table S3.

The highest likelihood was obtained for a model of speciation with gene flow, in which the rate of gene flow has recently increased (Fig. 3, Table S3). The change in the rate of gene flow was inferred to have occurred about 420 generations (~840 years) ago. However, these time estimates should be interpreted with caution as we fixed the first split to the upper bound of 6,000 generations ago (time since refilling of Lake Victoria after complete desiccation, Johnson, et al. 2000). Splitting times may be more recent if there was a significant time lag between refilling of the lake and the origin and beginning of diversification of *Pundamilia*. The second best model (ΔAIC 242) was the secondary contact model with only recent gene flow, estimated to have started 925 generations ago (~1,850 years). The models of only early gene flow and of constant gene flow had very similar likelihoods (ΔAIC 2,106 and 2,147, with overlapping likelihood distributions, Fig. 3). The model without gene flow had a very poor fit to the data (ΔAIC 12,710, Fig. 3). These results suggest that speciation occurred in the face of gene flow. However, having only two migration matrices (early and late) is a

strong simplification of the speciation process. Gene flow was likely very high when *P. nyererei* and *P. pundamilia* started to diverge and then gradually decreased as assortative mating and divergence in niche space increased, before it apparently increased again some 840 years ago. However, more individuals per species would be needed to model such additional complexity. With our current dataset, different models with more than two migration matrices cannot be significantly distinguished on the basis of their likelihoods (not shown).

Next, we tested if the patterns of genomic differentiation between the sympatric *Pundamilia* species at Makobe and at Python Islands and those between the allopatric Makobe and Python populations were most consistent with parallel, single, or hybrid origin models (Fig. 4, Fig. S2, Table S4). We compared four demographic models differing in topology: 1) a single speciation event between *P. nyererei* and *P. pundamilia* with subsequent independent colonization of Python and Makobe Islands by both species and interspecific gene flow in sympatry at each island but no gene flow between island populations of the same nominal species, 2) a parallel speciation model with two independent events of speciation into *P. nyererei* and *P. pundamilia*, wherein one species pair occurs at Makobe Island, and the other one at Python Island, allowing for subsequent gene flow between allopatric island populations of the same nominal species but none between sympatric species, 3) a hybrid origin model, wherein the red *Pundamilia* population from Python results from introgression between the blue *Pundamilia* population from Python and *P. nyererei* from Makobe Island, allowing for recent gene flow between the species in sympatry at both islands, and 4) a model of parallel speciation, whereby the Python species are derived from a hybrid ancestor, also allowing for recent gene flow in sympatry (Fig. 4a, Table S4). Given that our single island analysis suggested the occurrence of early gene flow (Fig. 3), we allowed for continuous gene flow between the two first populations after their initial split 6,000 generations ago. Among these models, the single speciation model (1) had the best model fit, followed by the model of speciation at Python from a hybrid population (model 4; Fig. 4a).

We also tested models without gene flow between the populations resulting from the initial split (early gene flow), allowing only for recent gene flow during the last 500 generations (Fig. S2). Consistent with the results from the models for Makobe Island that speciation occurred in the face of gene flow (Fig. 3), demographic models with only recent gene flow during the past 500 generations had low likelihoods and are thus not discussed in detail (Fig. S2). Additionally, we tested models of parallel speciation whereby the two species at Python Island form the sister clade either of *P. nyererei* or of *P. pundamilia* from Makobe Island (Fig. S2). Again, because of very low likelihoods of these models, we do not discuss them in depth.

Next, we tested if we could improve the fit with more realistic patterns of recent gene flow. We allowed for continuous migration between island populations of the same nominal species, and for continuous gene flow between the two species in sympatry (more realistic gene flow, Fig. 4, Fig. 5, Table S4) in addition to the early gene flow between the two populations arising from the first split 6,000 generations ago. Allowing for more realistic gene flow clearly improved the model fit, particularly for the hybrid origin models. The model of speciation from a hybrid population at Python is the best model among all the tested scenarios (Fig. 5, Fig. S2). Under this model, the first speciation event into *P. pundamilia* and *P. nyererei* sampled at Makobe is set to 6,000 generations ago. The inferred parameters suggest that about 4,800 generations later, *P. pundamilia* colonized the Mwanza Gulf (e.g. Python Island) and another 1,000 generations later *P. nyererei* also colonized the Mwanza Gulf and introgressed into the local *P. pundamilia* population (95% CI: 123-649 generations ago). Shortly after the introgression event (95% CI: 67-321 generations ago), the admixed population splits into two species (Fig. 5). This model suggests that the species sampled at Python Island evolved at least partially independently from those sampled at Makobe Island. We thus call the species at Python Island *P*. sp. “nyererei-like” (red) and *P.* sp. “pundamilia-like” (blue). The introgression event is estimated to have contributed a high proportion of the genetic variation

(47%) of both Python species. Recent gene flow was estimated to be especially high from the red species into the blue species at both islands and from *P*. sp. "nyererei-like" at Python to *P. nyererei* at Makobe Island (Fig. 5).

In order to better characterize the hybridization event, we tested alternative models of hybridization scenarios. Allowing for a second *P. nyererei* introgression event into the red Python species did not increase the likelihood and introgression was estimated to be very low (0.4%, Fig. S2). Furthermore, the model fit was lower if hybridization was modeled as having occurred in two separate events, one at the origin of each Python species, or if *P. nyererei* was the first species to colonize the Mwanza Gulf and speciation occurred from a hybrid population after *P. pundamilia* arrived in a second event and introgressed (Fig. S2).

Our parameter estimates could be biased if the four whole-genome sequenced individuals per population were not representative of their populations. As an example, if a *P.* sp. "nyererei-like" individual from Python Island was a recent hybrid, it would inflate the ancestry proportions and gene flow estimates. Therefore, we assessed the cluster proportions of the whole-genome-sequenced individuals compared to the RAD sequenced individuals (15-21 individuals per population) in a STRUCTURE analysis, and found that the ancestry proportions of the individuals used for whole-genome sequencing are representative of their respective populations (Fig. S3, S4). The STRUCTURE analysis further suggests that the red species at Python Island (*P.* sp. "nyererei-like") has a high similarity to *P. nyererei* Makobe and that the red and blue species at both islands in the Mwanza Gulf (Python and Kissenda) are highly similar.

Discussion

Discriminating between alternative speciation and migration scenarios to explain the current distribution of sister species in the genus *Pundamilia* is important for understanding the evolution of the Lake Victoria cichlid species flock. Here, we confirmed the findings by Seehausen, et al. (2008) that the red and blue *Pundamilia* cichlids are genomically distinct at all islands except for the highly turbid water Luanso Island, and that genetic differentiation between sympatric species decreases with water turbidity (Fig. 2). Second, we found that the speciation event into red and blue *Pundamilia* occurred in the face of gene flow (Fig. 3). Third, the *Pundamilia* individuals cluster by island and by species within island rather than by nominal species across islands (Fig. 2, Fig. S1). Demographic modeling shows that this is best explained by two separate origins of the species pairs at Makobe and Python Island, whereby both species at Python Island are derived from a hybrid population (Fig. 4+5). In the following we will discuss each of these findings and the general concepts and implications. Finally, we will propose a scenario for the evolution of the species complex by combining our modeling results with previous findings about mate choice and ecology of these species (Fig. 6).

Genetic clustering of individuals by species and geography

With the exception of Luanso Island, *P. nyererei*-like males with red nuptial coloration and *P. pundamilia*-like males with blue nuptial coloration belong to distinct, yet fully sympatric genetic clusters at each island, confirming earlier results (e.g. by Seehausen, et al. 2008; Seehausen 2009; Selz, et al. 2016) that they represent distinct species in sympatry at each island that largely mate assortatively. Some of the individuals from Python and Kissenda Island that were phenotypically intermediate were also genetically intermediate, but others were genetically indistinguishable from *P. nyererei* or from *P. pundamilia* individuals. This indicates that phenotypically intermediate individuals can represent backcross individuals, and are not all F1 or F2 hybrids.

Consistent with previous findings (Seehausen, et al. 2008), the genetic structure of the *Pundamilia* populations does not simply reflect the nominal species but reveals a more complex evolutionary history with gene flow. Among island populations of *P. pundamilia* we find overall increasing genetic distance with geographical distance between islands, consistent with isolation by distance (IBD). This is not the case in *P. nyererei*, consistent with earlier work using microsatellites (Seehausen, et al. 2008), and in line with evidence for more gene flow between islands in this species. The populations of each species at Kissenda and Python Islands in the northwestern Mwanza Gulf are each genetically very similar and cluster together on PC2 and PC3, suggesting that they share a similar evolutionary history. The same is supported also by the fact that they resemble each other phenotypically, *i.e.* both *P.* sp. "nyererei-like" populations belong share unique features morphology and of female and male coloration (Table S1, Seehausen 1996), and both *P.* sp. "pundamilia-like" populations have continuous broad vertical bars whereas *P. pundamilia* from Makobe Island has broken bars interrupted by a broken midlateral and dorsolateral stripe (Table S1, Seehausen 1996).

The larger genetic distance between populations from Python and Luanso Islands, despite smaller geographic distance in the central and southern Mwanza Gulf, respectively, may be due to genetic drift and interspecific gene flow at Luanso Island. Luanso individuals show no signs of genetic structure between male nuptial color morphs, nor do they show associations between nuptial coloration and multilocus (RADseq) genotype, suggesting that the Luanso individuals represent a single population with considerable color variation. It is possible that both *Pundamilia* species independently colonized Luanso Island and admixed completely in the turbid water. This would be consistent with previous findings that female mating preferences segregate in the Luanso population, with some females having significant preferences for either red or blue males whereas most females do not show color-based preferences (van der Sluijs, et al. 2008).

Speciation with gene flow

The comparison of alternative speciation models suggests that divergence between *P. nyererei* and *P. pundamilia* occurred in the face of gene flow. Due to the simplification of gene flow to one or two different migration matrices per model, we cannot completely rule out short periods of allopatry, but we consider this unlikely given the very wide geographical and ecological distribution of *P. pundamilia* and the patchy and fully sympatric distribution of *P. nyererei* nested within that of *P. pundamilia.*

We also find evidence for recent gene flow between species in sympatry as well as between islands within nominal species with demographic modeling (Fig. 4+5) and PCA (Fig. 2). Ongoing gene flow between sympatric species is not unexpected in a young adaptive radiation such as that of the Lake Victoria cichlids (Seehausen 2004; Grant and Grant 2008; Abbott, et al. 2013; Lamichhaney, et al. 2015). However, that the species remain distinct in full sympatry despite considerable interspecific gene flow implies a role for divergent or disruptive selection in sympatry (Dieckmann and Doebeli 1999; Kondrashov and Kondrashov 1999; Gavrilets 2004).

Hybrid origin of the Python Island species pair

Our demographic modeling approach suggests that speciation from a hybrid population in the Mwanza Gulf is the most likely model among all tested. If we constrain gene flow to occur among non-sister taxa populations (simple gene flow scenario), we find that the single speciation model reaches the highest relative likelihood (Fig. 4). However, our results suggest that there is considerable gene flow between allopatric populations of the same nominal species (see also Fig. 5). Indeed, if we include such migration events, the likelihood of the models increases considerably, with the speciation from a hybrid population model attaining the highest likelihood. Note that in the

simple gene flow scenarios we did not allow for migration between island populations, and that likely explains why the hybrid speciation model was not favoured in that case. Assuming that the speciation from a hybrid population model with migration among island populations is correct, by ignoring such migration rates the models are forced to reproduce the high genetic similarity between the *P. nyererei*-like island populations as recent shared ancestry. This finding highlights the importance of testing models with different complexities informed by observational data.

Asymmetric gene flow and effective population sizes

The higher amount of gene flow inferred to have occurred from *P. nyererei* into *P. pundamilia* than vice versa is consistent with the larger local census population sizes of *P. nyererei* (Seehausen 1996; Bouton, et al. 1997; Seehausen 1997; Seehausen and Bouton 1997). As in our models the amount of gene flow was not allowed to change over time, whereas in reality gene flow levels almost certainly were high early in speciation and then decreased as reproductive isolation strengthened, current gene flow may be lower than estimated in our models (or splitting times may be underestimated). The high gene flow estimates are supported by the presence of intermediate phenotypes at Python Island and also at Kissenda Island and are also predicted by the relatively turbid waters (Seehausen, et al. 2008). Yet, color-based assortative mating under clear water broad spectrum light conditions in the laboratory is strong among individuals collected from Python Island (Haesler and Seehausen 2005; Seehausen, et al. 2008; Selz, Pierotti, et al. 2014) and is likely important in retaining species differentiation despite some hybridization.

Allopatric gene flow between island populations likely did not occur directly from Makobe to Python or vice versa. Several islands between Makobe and Python Islands host *Pundamilia* populations and they are separated by unsuitable habitat (these species are rock specialists, and intervening habitat lacks rocks) (Fig. 1). Gene flow between the Makobe and Python Island population thus likely

occurred in an "island hopping" fashion, as is known from many studies of spatial genetic structure in rock specialist cichlids (Arnegard, et al. 1999; Wagner and McCune 2009; Koblmüller, et al. 2011).

The large ancestral effective population size inferred for *P. pundamilia* is consistent with the widespread occurrence of *P. pundamilia* in Lake Victoria and indicates that the ancestral population was structured. The much smaller ancestral population size of *P. nyererei* is in line with the patchy and much smaller distribution of that species (Fig. 1, Seehausen 1996; Seehausen 1997; Seehausen, et al. 1998; Seehausen 2009). However, at each island inhabited by both species, the local census population size of *P. nyererei* is generally larger than that of *P. pundamilia,* including at Python and Kissenda Islands, where *P.* sp. "nyererei-like" occupies the entire depth range, whereas *P.* sp. "pundamilia-like" is more restricted to shallow water (Seehausen 1996; Bouton, et al. 1997; Seehausen 1997; Seehausen and Bouton 1997). In the best fitting model this difference in local abundance is reflected in the larger population sizes of *P. nyererei* populations at both islands as compared to their sympatric *P. pundamilia* populations (Fig. 5).

"Hybrid parallel speciation" in Lake Victoria cichlids

A scenario consistent with our modeling results (Fig. 5, 6) and previous findings from research on these species is one where the original speciation event that produced *P. pundamilia* (blue) and *P. nyererei* (red) happened outside the Mwanza Gulf several thousand generations before the Mwanza Gulf was colonized by *P. pundamilia* but after the refilling of the lake 14,600 years ago. Eventually, about 1,000 generations after *P. pundamilia* had established populations in the Mwanza Gulf, massive introgression occurred from *P. nyererei* into the *P. pundamilia* population resulting in a hybrid population with similar ancestry proportions of *P. nyererei* and *P. pundamilia*. This may have happened through immigration of *P. nyererei* individuals from islands outside the Mwanza Gulf, which then hybridized with the local *P. pundamilia*.

Red coloration is dominant in laboratory crosses between the species (Magalhaes, et al. 2009). As the red coloration of initial *P. nyererei* immigrants and their F1 hybrid offspring would have made for more conspicuous visual signals than blue in deeper water (Seehausen, et al. 1997), and their red-shifted color-vision would have affected both mate preference and facilitated vision in deeper water (Maan, et al. 2006), *P. nyererei* alleles underlying these traits would have immediately had adaptive value in the Mwanza Gulf, and would probably have allowed the introgressed population of *Pundamilia* to expand into greater water depth and persisted in the deeper sections of the habitat. Consistent with this hypothesis, contemporary populations of *P.* sp. "nyererei-like" at Python and Kissenda Islands have nearly fixed the red-shifted *LWS* opsin allele (Seehausen, et al. 2008) and are most abundant at depths greater than 2 m whereas *P.* sp. "pundamilia-like" has the blue-shifted *LWS* allele and is most abundant at depths less than 2 m (Seehausen 2009). Assortative mating seems to be mainly due to divergent female preferences for red and blue male nuptial coloration (Stelkens, et al. 2008; Selz, Pierotti, et al. 2014) which have a relatively simple genetic basis (Haesler and Seehausen 2005). As male Lake Victoria cichlids often bias aggression towards males of their own color or the locally more common color phenotypes (Seehausen and Schluter 2004) and additionally *P. nyererei* males dominate *P. pundamilia* males in dyadic interactions (Dijkstra, et al. 2005), the first *P. nyererei* males migrating to islands in the Mwanza Gulf, and their red F1 hybrid offspring, may have had an advantage over the local *P. pundamilia* males in competition for territories, which would facilitate reproductive success. The interaction of sexual and natural selection acting on the same set of traits may subsequently have driven the evolution of a new species pair from within the admixture-enriched local *Pundamilia* population.

Even though we only tested the populations at Python and Makobe Islands, given that populations of the same color at Python and Kissenda Islands mostly cluster together in the PCA (Fig. 1), exhibit low F_{ST} (Fig. S1), and share the same phenotypic traits (Seehausen 1996), likely all *Pundamilia* populations at both islands and potentially at all islands in the north-western Mwanza Gulf (Fig. 1) are derived from the same hybrid speciation event. We note that our current data and analyses do

not allow us to clearly distinguish between a scenario whereby red and blue species re-emerged in the Mwanza Gulf after complete mixing of the ancestral *P. pundamilia* population with the late colonizing *P. nyererei,* or if some linkage disequilibrium among *P. nyererei* alleles and among *P. pundamilia* alleles was retained and facilitated rapid re-emergence of nyererei-like and pundamilia-like populations. Functional analyses and genome scans may help to discriminate between these alternatives.

To be considered a case of classical hybrid speciation, a hybrid population needs to evolve reproductive isolation from both parental lineages soon after the admixture event (Arnold 1997; Mallet 2007; Abbott, et al. 2010), and that does not appear to be the case in the situation studied here. On the other hand, parallel speciation as defined by Schluter and Nagel (1995) predicts that ecologically similar species which evolved in parallel lack reproductive isolation but show reproductive isolation to ecologically different species in sympatry. The *Pundamilia* species we studied here rather fulfill the expectations from parallel speciation with weaker reproductive isolation among allopatric species of similar phenotype and ecology than between sympatric species. *P.* sp. "nyererei-like" females at Python Island strongly prefer own-type males over sympatric *P.* sp. "pundamilia-like" males (Selz, Pierotti, et al. 2014) which closely resemble the *P. pundamilia* parental lineage. The traits crucial for assortative mating (red male coloration, preference for red, and red-shifted LWS alleles; Seehausen, et al. 2008; Selz, Pierotti, et al. 2014) are likely derived from the second parental lineage, *P. nyererei,* from outside the area (Fig. 6). Reproductive isolation between *P.* sp. "nyererei-like" at Python Island and the *P. nyererei* from Makobe Island is less straightforward, as both exhibit the same general type of male nuptial coloration, namely a bright red dorsum and dorsal fin and yellow flanks. However, as explained above, they are distinct in some other elements of nuptial coloration, such that the chroma of colour of different body and fin regions significantly differs between *P. nyererei* and *P. sp. "nyererei*-like" (Table S1, Seehausen 1996; Castillo Cajas, et al. 2012). There is also recent experimental evidence for behavioral reproductive isolation between *P.* sp. "nyererei-like" from Python and *P. nyererei* from Makobe (Selz, et al. 2016).

In the case of *P.* sp. "pundamilia-like", females prefer conspecific males over heterospecific *P.* sp. "nyererei-like males" (Selz, Pierotti, et al. 2014), but whether they also prefer *P.* sp. "pundamilia-like" males over *P. pundamilia* males awaits further testing.

However, *Pundamilia* are not a classical case of narrow-sense parallel speciation, because the evolutionary histories between the two red-blue species pairs are not completely independent (first criterion of Schluter and Nagel, 1995). This system may represent a special case of broad-sense parallel speciation initiated by introgressive hybridization between a pair of precursor species, which we term "hybrid parallel speciation" (Fig. 6). It is conceivable that the species from Python Island would merge with their allopatric parental populations of similar color (e.g. from Makobe Island) if they ever meet upon range expansion, much like the prediction for ecological parallel speciation (Schluter and Nagel 1995). Therefore, "hybrid parallel speciation" may not necessarily increase the total number of lasting biological species, but may often rather generate species with polyphyletic origins. If such parallel speciation events were common, the already exceptionally high speciation rate in Lake Victoria cichlids may have been yet underestimated.

In the "transporter hypothesis" (Schluter and Conte 2009) parallel evolution of a particular type of species is also facilitated by gene flow from a precursor species of similar type. This hypothesis was introduced to describe the parallel evolution of freshwater stickleback from marine stickleback in different freshwater catchments. Here, after the first origin of a freshwater-adapted stickleback population, freshwater-adapted alleles were introduced into the marine population through gene flow followed by multiple generations of recombination into marine genomic backgrounds. Upon colonization of a new river system from the Ocean, many freshwater alleles were recruited again through selection and freshwater genotypes are reassembled (Schluter and Conte 2009). We think that hybridization had a more direct effect in *Pundamilia*, as our high estimate of *P. nyererei* ancestry in the ancestral population of Python Island before speciation (47%, Fig. 5) suggests that a large fraction of the genomic variation introgressed from the parental *P. nyererei* lineage. This

increases the likelihood that linkage among parental alleles was retained to some extent in the hybrid population, which would facilitate rapid speciation in the absence of geographical isolation and might affect the rate at which the hybrid species stabilize and also the extent of parallelism with the original species pair.

Hybrid speciation has been hypothesized to underlie the origins of *Mbipia mbipi* and *Pundamilia* sp. "pink anal fin" from Lake Victoria (Keller, et al. 2013) and given that the same or very similar male nuptial color phenotypes (e.g. blue, red-back, red-belly) define sympatric sister species in most genera of Lake Victoria cichlids (Seehausen, et al. 1999), shuffling of the underlying alleles between non-sister species by occasional or geographically localized hybridization, followed by speciation, may have happened repeatedly. As the endemic species in Lake Victoria evolved in only 15,000 years (Stager and Johnson 2008), it is unlikely that the alleles underlying these color patterns arose repeatedly from *de novo* mutations, but rather form part of the standing genetic variation shared by many Lake Victoria cichlids. However, strong sexual selection is expected to deplete the standing variation at genes relevant to female preference and male nuptial coloration within each species (Seehausen 2004; Barrett and Schluter 2008). Standing genetic variation in key speciation traits may be recovered through occasional interspecific gene flow in adaptive radiations (syngameon hypothesis, Grant and Grant 1992; Seehausen 2004), and interspecific gene flow has been demonstrated in a large number of adaptive radiations including *Heliconius* butterflies (Gilbert 2003; Mavárez, et al. 2006; Jiggins, et al. 2008; Mallet 2009; The Heliconius Genome Consortium 2012), Darwin's finches (Grant and Grant 1994; Grant, et al. 2005; Lamichhaney, et al. 2015), *Mimulus* monkeyflowers (Stankowski and Streisfeld 2015), *Rhagoletis* fruit flies (Feder, et al. 2005), sunflowers (Rieseberg, et al. 2003), clownfishes (Litsios and Salamin 2014), sailfin silversides (Herder, et al. 2006), and cichlid fish of Lakes Tanganyika (Weiss, et al. 2015), Malawi (Genner and Turner 2012), Victoria (Keller, et al. 2013) and Barombi Mbo (Schliewen and Klee 2004). More work is needed to make a more direct link between interspecific gene flow and diversification in Lake Victoria cichlids. Importantly, identifying the genomic regions under divergent selection or involved

in reproductive isolation between species would allow assessment of the relative importance of introgression as a source of genomic variation of relevance to speciation.

Conclusions

Here, we show how demographic modeling can improve our understanding of the processes underlying the observed patterns of genetic similarity and dissimilarity in a complex of closely related cichlid species from Lake Victoria. We provide evidence that similar *Pundamilia* sister species pairs at different islands have arisen twice, where the second event is associated with introgressive hybridization between a pair of precursor species. Our study is consistent with a role of interspecific gene flow in the diversification of Lake Victoria cichlids, as proposed in the syngameon hypothesis for adaptive radiations. Further investigations should target the key mechanisms in the diversification of *Pundamilia*, the genes and genomic architecture underlying the relevant traits and their evolutionary histories.

Acknowledgements

We thank the Tanzanian Commission for Science and Technology for research permission, and the Tanzanian Fisheries Research Institute for facilities and logistical support. We are indebted to Mhoja Kayeba and Mohammed Haluna for technical assistance in the field and to Inke van der Sluijs and Nellie Konijnendijk for collecting some of the tissue samples used in this study. We also thank Marcel Häsler and Salome Mwaiko of the Department of Fish Ecology and Evolution at EAWAG, Keith Harshman of the Lausanne Genomic Technologies Facility, and Cord Drögemüller, Tosso Leeb, Muriel Fragnière, and Michèle Ackermann of the NGS platform of the University of Bern for Illumina sequencing and lab support. We thank Stefan Zoller from the Genetic Diversity Center (GDC) at ETH Zürich for bioinformatics support. Genomic analyses were performed using the computing

infrastructure of the GDC, the Euler computer cluster at ETH Zurich, and the Ubelix computer cluster at University of Bern, Switzerland. This research was supported by the Swiss National Science Foundation grant PDFMP3 134657 to OS and LE.

References

Abbott R, Albach D, Ansell S, Arntzen JW, Baird SJ, Bierne N, Boughman J, Brelsford A, Buerkle CA, Buggs R, et al. 2013. Hybridization and speciation. J Evol Biol 26:229-246.

Abbott RJ, Hegarty MJ, Hiscock SJ, Brennan AC. 2010. Homoploid hybrid speciation in action. Taxon 59:1375-1386.

Akaike H. 1974. A new look at the statistical model identification. IEEE Trans Autom Control 19:716-723.

Allender CJ, Seehausen O, Knight ME, Turner GF, Maclean N. (Allender2003 co-authors). 2003. Divergent selection during speciation of Lake Malawi cichlid fishes inferred from parallel radiations in nuptial coloration. Proc Natl Acad Sci USA 100.

Arendt J, Reznick D. 2008. Convergence and parallelism reconsidered: what have we learned about the genetics of adaptation? Trends Ecol Evol 23:26-32.

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

Arnold ML. 1997. Natural hybridization and evolution: Oxford University Press.

Baird NA, Etter PD, Atwood TS, Currey MC, Shiver AL, Lewis ZA, Selker EU, Cresko WA, Johnson EA. 2008. Rapid SNP Discovery and Genetic Mapping Using Sequenced RAD Markers. Plos One 3:e3376.

Barrett RDH, Schluter D. 2008. Adaptation from standing genetic variation. Trends Ecol Evol 23:38-44.

Barton N, Hewitt GM. 1989. Adaptation, speciation and hybrid zones. Nature 341:497-503.

Barton NH, Hewitt GM. 1985. Analysis of hybrid zones. Ann Rev Ecol Syst:113-148.

Bierne N, Gagnaire P-A, David P. 2013. The geography of introgression in a patchy environment and the thorn in the side of ecological speciation. Curr Zool 59:72-86.

Bouton N, Seehausen O, van Alphen JJM. 1997. Resource partitioning among rock-dwelling haplochromines (Pisces : Cichlidae) from Lake Victoria. Ecology of Freshwater Fish 6:225-240.

Brawand D, Wagner CE, Li YI, Malinsky M, Keller I, Fan SH, Simakov O, Ng AY, Lim ZW, Bezault E, et al. 2014. The genomic substrate for adaptive radiation in African cichlid fish. Nature 513:375-381.

Butlin RK, Saura M, Charrier G, Jackson B, André C, Caballero A, Coyne JA, Galindo J, Grahame JW, Hollander J. 2014. Parallel evolution of local adaptation and reproductive isolation in the face of gene flow. Evolution 68:935-949.

Canty A, Ripley B. 2015. R Package 'boot' (url: https://cran.r-project.org/web/packages/boot/). In.

Castillo Cajas RF, Selz OM, Ripmeester EA, Seehausen O, Maan ME. 2012. Species-specific relationships between water transparency and male coloration within and between two closely related Lake Victoria cichlid species. Int J Evol Biol 2012.

Catchen J, Hohenlohe PA, Bassham S, Amores A, Cresko WA. 2013. Stacks: an analysis tool set for population genomics. Mol Ecol 22:3124-3140.

Conte GL, Arnegard ME, Peichel CL, Schluter D. 2012. The probability of genetic parallelism and

convergence in natural populations. Proc Biol Sci B 279:5039-5047.
Coyne JA, Orr HA. 2004. Speciation: Sinauer Associates Sunderland, MA.
Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, et al. 2011. The variant call format and VCFtools. Bioinformatics 27:2156-2158.
Dieckmann U, Doebeli M. 1999. On the origin of species by sympatric speciation. Nature 400:354-357.
Dijkstra PD, Seehausen O, Groothuis TG. 2005. Direct male-male competition can facilitate invasion of new colour types in Lake Victoria cichlids. Behav Ecol Sociobiol 58:136-143.
Dijkstra PD, Seehausen O, Pierotti MER, Groothuis TGG. 2007. Male-male competition and speciation: aggression bias towards differently coloured rivals varies between stages of speciation in a Lake Victoria cichlid species complex. J Evol Biol 20:496-502.
Earl D, vonHoldt B. 2012. STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. Conserv Biol Res 4:359-361.
Evanno G, Regnaut S, Goudet J. 2005. Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. Mol Ecol 14:2611-2620.
Excoffier L, Dupanloup I, Huerta-Sánchez E, Sousa VC, Foll M. 2013. Robust Demographic Inference from Genomic and SNP Data. PLoS Genet 9:e1003905.
Excoffier L, Lischer HEL. 2010. Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. Mol Ecol Res 10:564-567.
Faria R, Renaut S, Galindo J, Pinho C, Melo-Ferreira J, Melo M, Jones F, Salzburger W, Schluter D, Butlin R. 2014. Advances in ecological speciation: an integrative approach. Mol Ecol 23:513-521.
Feder JL, Xie X, Rull J, Velez S, Forbes A, Leung B, Dambroski H, Filchak KE, Aluja M. 2005. Mayr, Dobzhansky, and Bush and the complexities of sympatric speciation in Rhagoletis. Proc Natl Acad Sci USA 102:6573-6580.
Gavrilets S. 2004. Fitness landscapes and the origin of species (MPB-41): Princeton University Press Princeton, NJ.
Genner MJ, Turner GF. 2012. Ancient Hybridization and Phenotypic Novelty within Lake Malawi's Cichlid Fish Radiation. Mol Biol Evol 29:195-206.
Gilbert L. 2003. Adaptive novelty through introgression in Heliconius wing patterns: evidence for shared genetic “tool box” from synthetic hybrid zones and a theory of diversification. Ecology and Evolution Taking Flight: Butterflies as Model Systems:281-318.
Gompert Z, Buerkle CA. 2009. A powerful regression-based method for admixture mapping of isolation across the genome of hybrids. Mol Ecol 18:1207-1224.
Grant BR, Grant PR. 2008. Fission and fusion of Darwin's finches populations. Phil Trans R Soc B 363:2821-2829.
Grant PR, Grant BR. 1992. Hybridization of Bird Species. Science 256:193-197.
Grant PR, Grant BR. 1994. Phenotypic and genetic effects of hybridization in Darwin's finches. Evolution:297-316.
Grant PR, Grant BR, Petren K. 2005. Hybridization in the recent past. Am Nat 166.
Haas O, Simpson GG. 1946. Analysis of some phylogenetic terms, with attempts at redefinition. Proc Am Philos Soc:319-349.
Haesler MP, Seehausen O. 2005. Inheritance of female mating preference in a sympatric sibling species pair of Lake Victoria cichlids: implications for speciation. Proc Biol Sci B 272:237-245.
Harrison RG, Larson EL. 2016. Heterogeneous genome divergence, differential introgression, and the origin and structure of hybrid zones. Mol Ecol.
Herder F, Nolte AW, Pfaender J, Schwarzer J, Hadiaty RK, Schliewen UK. 2006. Adaptive radiation and hybridization in Wallace's Dreamponds: evidence from sailfin silversides in the Malili Lakes of Sulawesi. Proc Biol Sci B 273:2209-2217.
Jakobsson M, Rosenberg NA. 2007. CLUMPP: a cluster matching and permutation program for

dealing with label switching and multimodality in analysis of population structure. Bioinformatics 23:1801-1806.
Jiggins CD, Salazar C, Linares M, Mavarez J. 2008. Hybrid trait speciation and Heliconius butterflies. Philos Trans R Soc Lond B Biol Sci 363:3047-3054.
Johannesson K, Panova M, Kemppainen P, André C, Rolan-Alvarez E, Butlin RK. 2010. Repeated evolution of reproductive isolation in a marine snail: unveiling mechanisms of speciation. Philos Trans R Soc Lond B Biol Sci 365:1735-1747.
Johnson TC, Kelts K, Odada E. 2000. The holocene history of Lake Victoria. Ambio 29:2-11.
Jones FC, Grabherr MG, Chan YF, Russell P, Mauceli E, Johnson J, Swofford R, Pirun M, Zody MC, White S. 2012. The genomic basis of adaptive evolution in threespine sticklebacks. Nature 484:55-61.
Keller I, Wagner CE, Greuter L, Mwaiko S, Selz OM, Sivasundar A, Wittwer S, Seehausen O. 2013. Population genomic signatures of divergent adaptation, gene flow and hybrid speciation in the rapid radiation of Lake Victoria cichlid fishes. Mol Ecol 22:2848-2863.
Koblmüller S, Salzburger W, ObermÜLler B, Eigner EVA, Sturmbauer C, Sefc KM. 2011. Separated by sand, fused by dropping water: habitat barriers and fluctuating water levels steer the evolution of rock-dwelling cichlid populations in Lake Tanganyika. Mol Ecol 20:2272-2290.
Kondrashov AS, Kondrashov FA. 1999. Interactions among quantitative traits in the course of sympatric speciation. Nature 400:351-354.
Konijnendijk N, Joyce DA, Mrosso HDJ, Egas M, Seehausen O. 2011. Community Genetics Reveal Elevated Levels of Sympatric Gene Flow among Morphologically Similar but Not among Morphologically Dissimilar Species of Lake Victoria Cichlid Fish. Int J Evol Biol 2011:616320.
Kozarewa I, Ning Z, Quail MA, Sanders MJ, Berriman M, Turner DJ. 2009. Amplification-free Illumina sequencing-library preparation facilitates improved mapping and assembly of GC-biased genomes. Nat Methods 6:291-295.
Lamichhaney S, Berglund J, Almén MS, Maqbool K, Grabherr M, Martinez-Barrio A, Promerová M, Rubin C-J, Wang C, Zamani N. 2015. Evolution of Darwin/'s finches and their beaks revealed by genome sequencing. Nature 518:371-375.
Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. Nat Methods 9:357-U354.
Litsios G, Salamin N. 2014. Hybridisation and diversification in the adaptive radiation of clownfishes. BMC Evol Biol 14:1-9.
Maan ME, Hofker KD, van Alphen JJM, Seehausen O. 2006. Sensory drive in cichlid speciation. Am Nat 167:947-954.
Maan ME, Seehausen O, Soderberg L, Johnson L, Ripmeester EAP, Mrosso HDJ, Taylor MI, van Dooren TJM, van Alphen JJM. 2004. Intraspecific sexual selection on a speciation trait, male coloration, in the Lake Victoria cichlid Pundamilia nyererei. Proc Biol Sci B 271:2445-2452.
Maan ME, Seehausen O, Van Alphen JJM. 2010. Female mating preferences and male coloration covary with water transparency in a Lake Victoria cichlid fish. Biol J Linn Soc 99:398-406.
Magalhaes I, Lundsgaard-Hansen B, Mwaiko S, Seehausen O. 2012. Evolutionary divergence in replicate pairs of ecotypes of Lake Victoria cichlid fish. Evol Ecol Res 14:381-401.
Magalhaes IS, Mwaiko S, Schneider MV, Seehausen O. 2009. Divergent selection and phenotypic plasticity during incipient speciation in Lake Victoria cichlid fish. J Evol Biol 22:260-274.
Mallet J. 2007. Hybrid speciation. Nature 446:279-283.
Mallet J. 2009. Rapid speciation, hybridization and adaptive radiation in the Heliconius melpomene group. Speciation and patterns of diversity:177-194.
Marques DA, Lucek K, Meier JI, Mwaiko S, Wagner CE, Excoffier L, Seehausen O. 2016. Genomics of Rapid Incipient Speciation in Sympatric Threespine Stickleback. PLoS Genet 12:e1005887.
Mavárez J, Salazar CA, Bermingham E, Salcedo C, Jiggins CD, Linares M. 2006. Speciation by hybridization in Heliconius butterflies. Nature 441:868-871.
McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D,

Gabriel S, Daly M, et al. 2010. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 20:1297-1303.
Nosil P. 2012. Ecological Speciation. Oxford: Oxford University Press.
Nosil P, Crespi BJ, Sandoval CP. 2002. Host-plant adaptation drives the parallel evolution of reproductive isolation. Nature 417:440-443.
Nosil P, Feder JL. 2013. Genome evolution and speciation: toward quantitative descriptions of pattern and process. Evolution 67:2461-2467.
Nosil P, Schluter D. 2011. The genes underlying the process of speciation. Trends Ecol Evol 26:160-167.
Okullo W, Ssenyonga T, Hamre B, Frette Ø, Sørensen K, Stamnes JJ, Steigen A, Stamnes K. 2007. Parameterization of the inherent optical properties of Murchison Bay, Lake Victoria. Applied optics 46:8553-8561.
Pritchard JK, Stephens M, Donnelly P. (Pritchard2000 co-authors). 2000. Inference of population structure using multilocus genotype data. Genetics 155.
Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ, et al. 2007. PLINK: A tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 81:559-575.
Rand DM, Harrison RG. 1989. Ecological genetics of a mosaic hybrid zone: mitochondrial, nuclear, and reproductive differentiation of crickets by soil type. Evolution:432-449.
Ravinet M, Westram A, Johannesson K, Butlin R, André C, Panova M. 2015. Shared and nonshared genomic divergence in parallel ecotypes of Littorina saxatilis at a local scale. Mol Ecol. 25: 287-305
Renaut S, Owens GL, Rieseberg LH. 2014. Shared selective pressure and local genomic landscape lead to repeatable patterns of genomic divergence in sunflowers. Mol Ecol 23:311-324.
Rieseberg LH, Buerkle CA. 2002. Genetic mapping in hybrid zones. Am Nat 159:S36-S50.
Rieseberg LH, Raymond O, Rosenthal DM, Lai Z, Livingstone K, Nakazato T, Durphy JL, Schwarzbach AE, Donovan LA, Lexer C. 2003. Major ecological transitions in wild sunflowers facilitated by hybridization. Science 301:1211-1216.
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.
Rosenblum EB, Parent CE, Brandt EE. 2014. The molecular basis of phenotypic convergence. Annu Rev Ecol Evol Syst 45:203-226.
Rundle HD, Nagel L, Boughman JW, Schluter D. 2000. Natural selection and parallel speciation in sympatric sticklebacks. Science 287:306-308.
Sambrook J, Russell DW. 2001. Molecular cloning : a laboratory manual. Cold Spring Harbor, N.Y.: Cold Spring Harbor Laboratory Press.
Schliewen UK, Klee B. 2004. Reticulate sympatric speciation in Cameroonian crater lake cichlids. Front Zool 1.
Schluter D, Clifford EA, Nemethy M, McKinnon JS. 2004. Parallel evolution and inheritance of quantitative traits. the american naturalist 163:809-822.
Schluter D, Conte GL. 2009. Genetics and ecological speciation. Proc Natl Acad Sci USA 106:9955-9962.
Schluter D, Nagel LM. 1995. Parallel Speciation by Natural Selection. Am Nat 146:292-301.
Seehausen O. 2006. African cichlid fish: a model system in adaptive radiation research. Proc R Soc B 273:1987-1998.
Seehausen O. 1997. Distribution of and reproductive isolation among color morphs of a rock-dwelling Lake Victoria cichlid (*Haplochromis nyererei*). Ecology of Freshwater Fish 6:59-66.
Seehausen O. 2004. Hybridization and adaptive radiation. Trends Ecol Evol 19:198-207.
Seehausen O. 1996. Lake Victoria Rock Cichlids. Taxonomy, Ecology and Distribution.
Seehausen O. 2015. Process and pattern in cichlid radiations–inferences for understanding unusually high rates of evolutionary diversification. New Phytol 207:304-312.

Seehausen O. 2009. Progressive levels of trait divergence along a “speciation transect” in the Lake Victoria cichlid fish Pundamilia. In: Butlin RK, Bridle J, Schluter D, editors. Speciation and patterns of diversity. Cambridge: Cambridge University Press. p. 155-176.

Seehausen O, Bouton N. 1997. Microdistribution and fluctuations in niche overlap in a rocky shore cichlid community in Lake Victoria. Ecology of Freshwater Fish 6:161-173.

Seehausen O, Butlin RK, Keller I, Wagner CE, Boughman JW, Hohenlohe PA, Peichel CL, Saetre G-P, Bank C, Brännström Å. 2014. Genomics and the origin of species. Nat Rev Genet 15:176-192.

Seehausen O, Lippitsch E, Bouton N, Zwennes H. 1998. Mbipi, the rock-dwelling cichlids of Lake Victoria: description of three new genera and fifteen new species. Ichthyol Explor Freshw 9:129-228.

Seehausen O, Schluter D. 2004. Male-male competition and nuptial-colour displacement as a diversifying force in Lake Victoria cichlid fishes. Proc Biol Sci B 271:1345-1353.

Seehausen O, Terai Y, Magalhaes IS, Carleton KL, Mrosso HDJ, Miyagi R, van der Sluijs I, Schneider MV, Maan ME, Tachida H, et al. 2008. Speciation through sensory drive in cichlid fish. Nature 455:620-U623.

Seehausen O, van Alphen J, Witte F. 1999. Can ancient colour polymorphisms explain why some cichlid lineages speciate rapidly under disruptive sexual selection? Belg J Zool 129:43-60.

Seehausen O, van Alphen JJM. 1998. The effect of male coloration on female mate choice in closely related Lake Victoria cichlids (Haplochromis nyererei complex). Behav Ecol Sociobiol 42:1-8.

Seehausen O, van Alphen JJM, Witte F. 1997. Cichlid fish diversity threatened by eutrophication that curbs sexual selection. Science 277:1808-1811.

Seehausen O, van Alphen JM. 1999. Can sympatric speciation by disruptive sexual selection explain rapid evolution of cichlid diversity in Lake Victoria? Ecol Lett 2:262-271.

Selz O, Lucek K, Young K, Seehausen O. 2014. Relaxed trait covariance in interspecific cichlid hybrids predicts morphological diversity in adaptive radiations. J Evol Biol 27:11-24.

Selz OM, Pierotti MER, Maan ME, Schmid C, Seehausen O. 2014. Female preference for male color is necessary and sufficient for assortative mating in 2 cichlid sister species. Behav Ecol 25:612-626.

Selz OM, Thommen R, Maan ME, Seehausen O. 2014. Behavioural isolation may facilitate homoploid hybrid speciation in cichlid fish. J Evol Biol 27:275-289.

Selz OM, Thommen R, Pierotti MER, Anaya-Rojas JM, Seehausen O. 2016. Differences in male coloration are predicted by divergent sexual selection between populations of a cichlid fish. Proc R Soc B 283:20160172

Stager JC, Johnson TC. 2008. The late Pleistocene desiccation of Lake Victoria and the origin of its endemic biota. Hydrobiologia 596:5-16.

Stankowski S, Streisfeld MA. 2015. Introgressive hybridization facilitates adaptive divergence in a recent radiation of monkeyflowers. Proc Biol Sci B 282.

Stelkens RB, Pierotti MER, Joyce DA, Smith AM, van der Sluijs I, Seehausen O. 2008. Disruptive sexual selection on male nuptial coloration in an experimental hybrid population of cichlid fish. Proc Biol Sci B 363:2861-2870.

Stern DL, Orgogozo V. 2009. Is genetic evolution predictable? Science 323:746-751.

Teeter KC, Thibodeau LM, Gompert Z, Buerkle CA, Nachman MW, Tucker PK. 2010. The variable genomic architecture of isolation between hybridizing species of house mice. Evolution 64:472-485.

The Heliconius Genome Consortium. 2012. Butterfly genome reveals promiscuous exchange of mimicry adaptations among species. Nature 487:94-98.

van der Sluijs I, van Alphen JJM, Seehausen O. 2008. Preference polymorphism for coloration but no speciation in a population of Lake Victoria cichlids. Behav Evol 19:177-183.

Wagner CE, McCune AR. 2009. Contrasting patterns of spatial genetic structure in sympatric rock-dwelling cichlid fishes. Evolution 63:1312-1326.

Weiss JD, Cotterill FP, Schliewen UK. 2015. Lake Tanganyika—A'Melting Pot'of Ancient and Young

Cichlid Lineages (Teleostei: Cichlidae)? Plos One 10:e0125043.
Zheng XW, Levine D, Shen J, Gogarten SM, Laurie C, Weir BS. 2012. A high-performance computing toolset for relatedness and principal component analysis of SNP data. Bioinformatics 28:3326-3328.

Data Accessibility

Demultiplexed fastq files of RADseq and whole-genome datasets have been deposited at NCBI SRA under BioProject PRJNA339828 (http://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA339828) and BioSamples SAMN05711156- SAMN05711318. Input files used for fastsimcoal2 and observed site frequency spectra are available on Dryad under doi:10.5061/dryad.2jd45.

Author Contributions

OS and JIM designed the study; OMS collected, and OS and OMS identified the cichlid samples; JIM performed the lab work and conducted the analyses, with assistance from OS, VCS, LE, DAM, and CEW. JIM wrote the manuscript with contributions from OS, LE, CEW, DAM, VCS, and OMS.

Tables and Figures

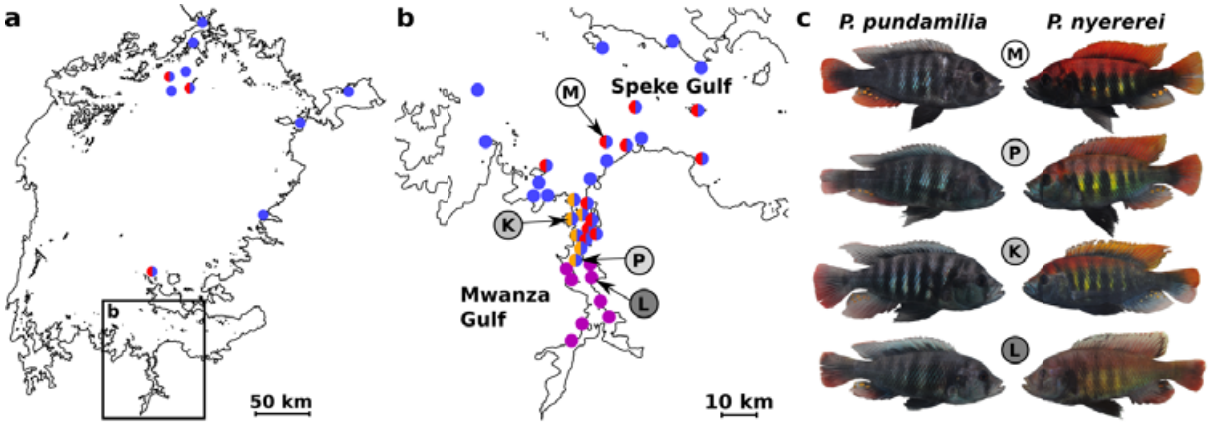

Figure 1. Distribution and sampling sites of *Pundamilia pundamilia* and *P. nyererei*. a. Map of Lake Victoria showing the known records of *P. pundamilia* (blue), *P. nyererei* (red) and sites harboring both species (half red, half blue circles). The square indicates the location of the Mwanza Gulf and Speke Gulf enlarged in **b**. The two types of *P. nyererei*, deep-bodied with blueish belly and slender-bodied without blueish belly (see Table S1), are marked with orange and dark red, respectively. Sites in the turbid South of the Mwanza Gulf inhabited by color-polymorphic *Pundamilia* populations which are not clearly identifiable to a species are shown with purple circles. Sampling sites of fish used in this study are indicated with single letter abbreviations (M=Makobe Island, K=Kissenda Island, P=Python Island, L=Luanso Island). The darkness of the grey background in the symbols reflects the water turbidity (darker=more turbid). Note that although Kissenda Island is closer to the open lake than Python Island, the former is more turbid as it lies inside a bay. Distribution records are assembled from (Seehausen 1996, 1997; Seehausen, et al. 1998; Seehausen 2009). **c.** Representative individuals of blue and red *Pundamilia* found at the different sampling sites. At Makobe, Kissenda and Python Islands, they represent the nominal species of *P. pundamilia* and *P. nyererei*, whereas for Luanso Island, red and blue individuals were selected from the polymorphic population.

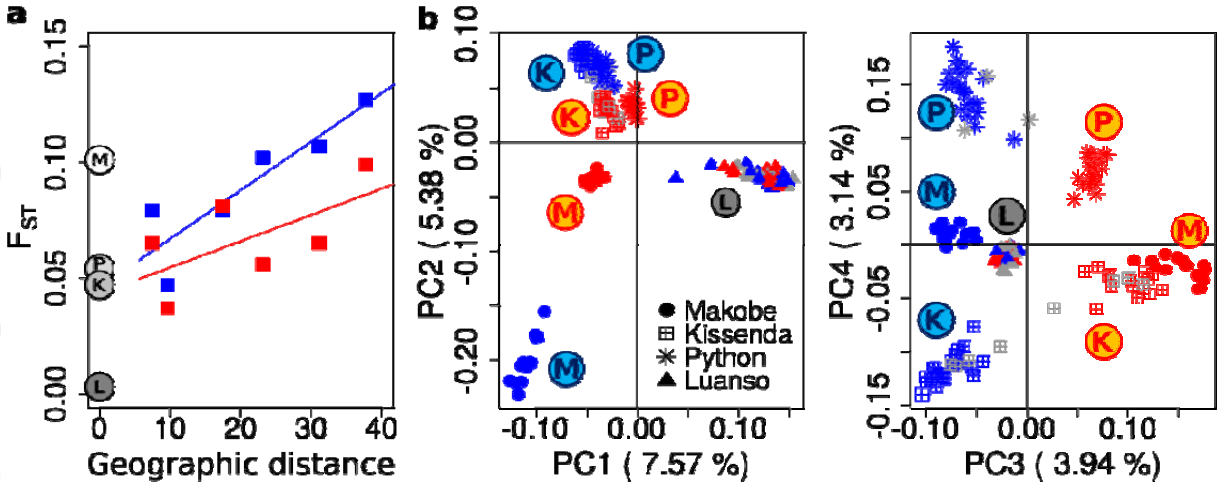

Figure 2. The major axis of genetic diversity does not separate the nominal species but shows an isolation by distance pattern. a Correlation of geographic distance with pairwise F_{ST} values between red populations (nyererei-like, red squares, not significant) and between blue populations (*pundamilia*-like, blue squares, Mantel test p-value=0.048) from different islands. Interspecific F_{ST} values between species in sympatry are shown as grey circles for each island at zero geographic distance (M=Makobe, P=Python, K=Kissenda, L=Luanso). As the *Pundamilia* at Luanso Island cannot be clearly assigned to a nominal species, red and blue individuals were identified by eye and used to estimate F_{ST}. The estimates are based on 15,017 biallelic sites (for pairwise F_{ST} comparisons between all populations, see Fig. S1). **b** Principal components analysis (PCA) showing the genetic similarity between the sampled individuals based on 13,213 biallelic sites from the RADseq dataset. The first four axes are shown with percentage of variance explained in parentheses. Different symbols represent individuals from different sampling sites as indicated in the legend. For Makobe, Python and Kissenda Islands, the red, blue or grey color indicates an individual is phenotypically identified as *P. nyererei*, *P. pundamilia*, or as a potential hybrid of intermediate phenotype, respectively. The Luanso Island individuals with most red or least red are shown with red or blue triangles, respectively, and others with grey triangles.

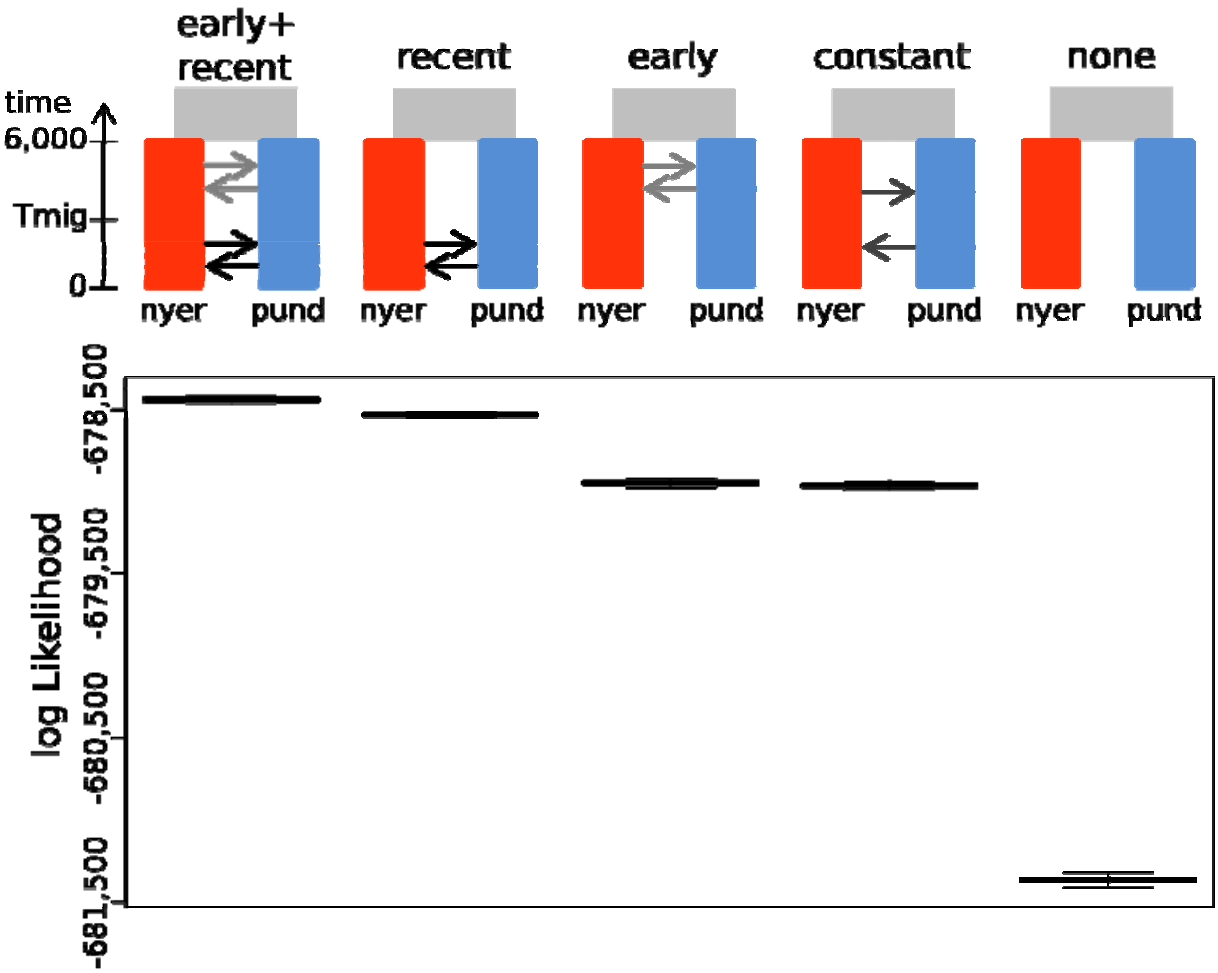

Figure 3. Comparing alternative gene flow scenarios reveals that a model of speciation with two different migration periods fits the data best. Tested models of the evolution of *P. nyererei* (nyer) and *P. pundamilia* (pund) at Makobe Island ordered by likelihood with boxplots showing the $\log_{10}$ likelihood distributions inferred by recalculating the $\log_{10}$ likelihoods for 100 simulated SFS. The first model (early+recent) allows for two periods with different migration rates (different early and recent gene flow matrices). The time of migration matrix change is inferred as a model parameter (Tmig). The second best model (recent) allows only for gene flow starting at Tmig until the present. The early gene flow model allows only for gene flow up to time Tmig and the constant gene flow model simulates ongoing migration that remains constant since the speciation time. The last model does not allow for gene flow.

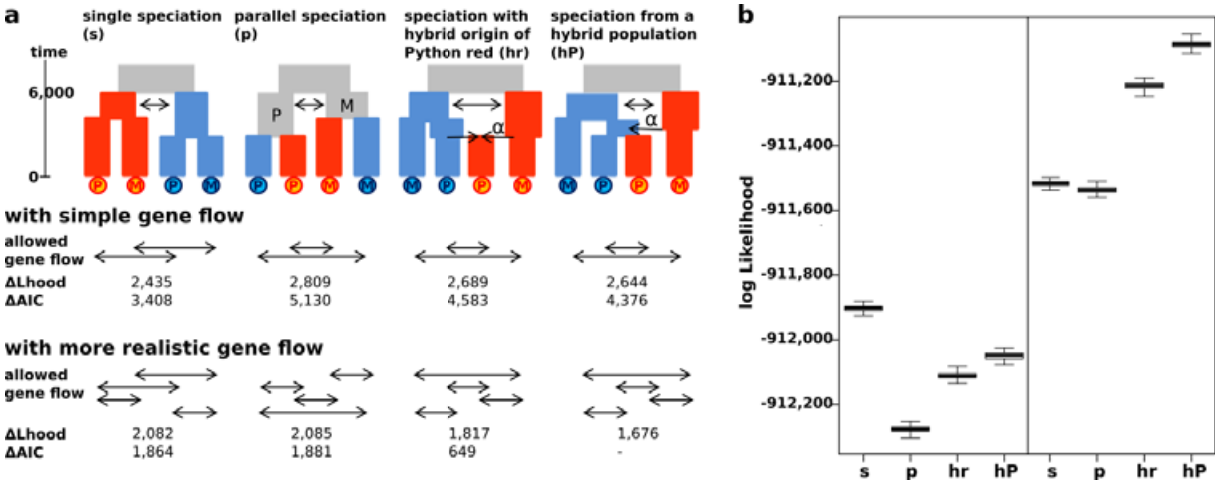

Figure 4. Comparison of demographic models reveals highest likelihood for a model with speciation from a hybrid ancestor. a Illustrations of demographic models with different tree topologies (single, parallel, and hybrid origin of one or both Python populations) with different gene flow scenarios (simple or more realistic). The numbers below the models denote the delta likelihoods (ΔLhood, difference between maximum possible and obtained model likelihood in $\log_{10}$ units) and delta AIC (ΔAIC) to the overall best model (speciation from a hybrid population at Python). **b** Boxplots showing the log Likelihood distributions from 100 expected SFS each approximated using 1 million coalescent simulations under the parameters that maximize the likelihood for each model. For each gene flow category (simple and more realistic), the likelihoods are given for the models of single speciation (s), parallel speciation (p), speciation with hybrid origin of the Python red population (hr), and speciation from a hybrid population (hP). Other models tested and associated likelihoods are given in Fig. S2.

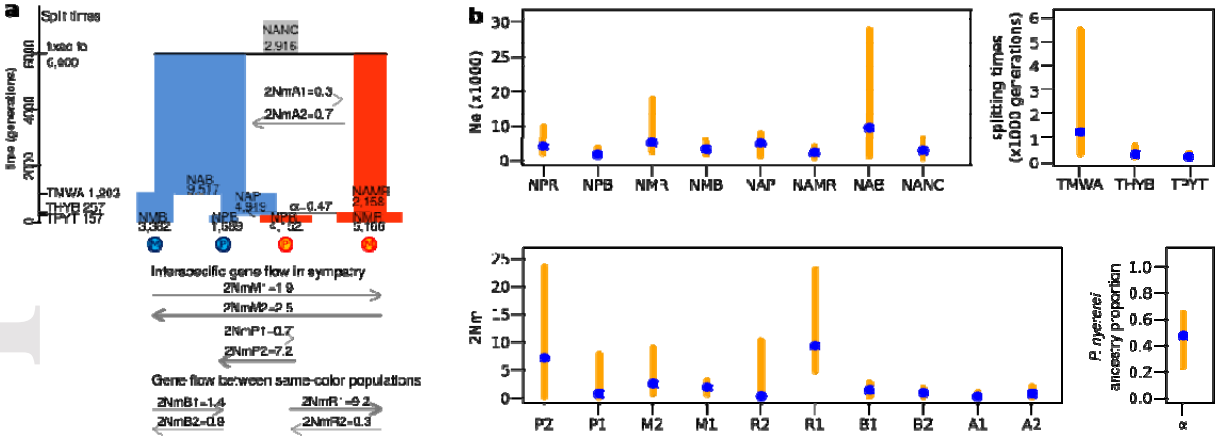

Figure 5. The demographic model with the best fit to the observed data is a model with speciation from a hybrid ancestor giving rise to the species pair at Python Island. a. Model illustration with parameter point estimates. Rectangles represent populations, whereas the width indicates effective population size (N) and the height corresponds to the inferred splitting times (T). THYB refers to the time of the *P. nyererei* introgression event which is shortly followed (within about 100 generations) by the split into two species at Python island at TPYT. Ancestral population sizes are labeled as "NA", whereas "NANC" refers to the size of the population ancestral to all sampled populations. Islands are abbreviated as "P" and "M" for Python and Makobe, respectively, and the nominal species as "B" for *P. pundamilia* (blue) and "R" for *P. nyererei* (red). The *P. nyererei* introgression event into the Python ancestor is indicated with a black arrow with ancestry proportion α .Grey arrows indicate gene flow given as effective number of haploid immigrants per generation (2Nm) forward in time. **b.** Confidence intervals for parameter estimates based on 200 bootstrap replicates. The point estimates are shown with blue dots and the 95% confidence intervals with orange vertical bars. For parameter names, see illustration in a.

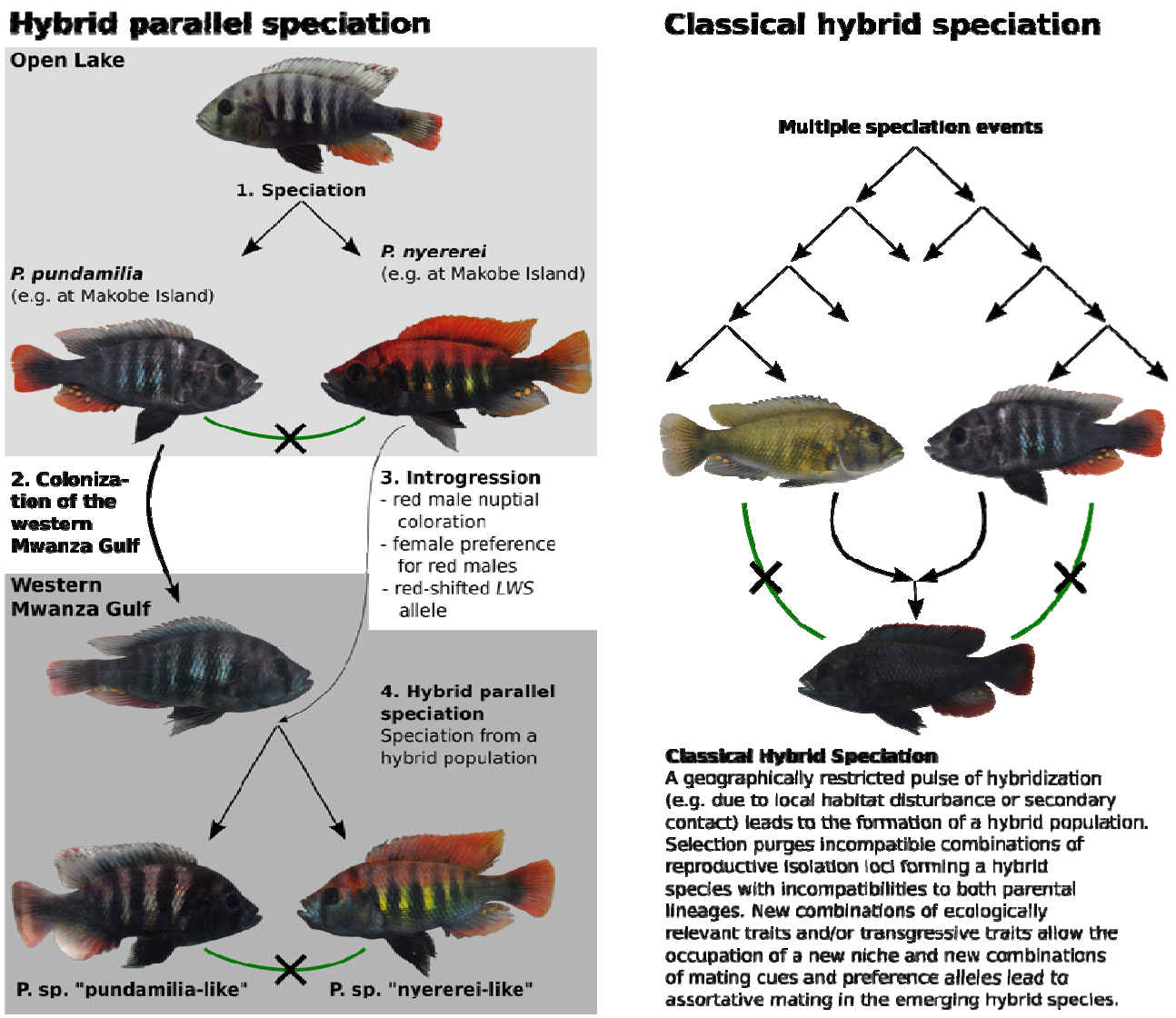

Figure 6. Illustration of the proposed evolutionary scenario of hybrid parallel speciation and a comparison to classical hybrid speciation. A scenario consistent with our results is one of initial speciation from a blue ancestor into *P. pundamilia* and *P. nyererei* (here represented by the populations at Makobe Island, M), followed much later by the colonization of the Mwanza Gulf (including Python Island, P) by *P. pundamilia*. Later again *P. nyererei* arrive in the Mwanza Gulf, hybridize with the local *P. pundamilia* population and merge into a hybrid population that carries a mixture of genetic variation from both species. This hybrid population speciates again into sympatric *P.* sp. *"pundamilia*-like" and *P.* sp. "nyererei-like". There is strong assortative mating between *P.* sp. "pundamilia-like" and *P.* sp. "nyererei-like" at Python (illustrated by the crossed green lines, Selz, Pierotti, et al. 2014) and also some evidence for assortative mating between individuals of *P.* sp. "nyererei-like" from Python and individuals of the original *P. nyererei* from Makobe Island (Selz et al., 2016). "Hybrid parallel speciation" relies on the local availability of the ecological niches of both the local and the geographically distant parental species giving hybrids with alternative trait combinations fitness advantages in different parts of the ecological range of the hybrid population. Should the geographical ranges of *P. sp. "nyererei*-like" and *P. nyererei* ever expand sufficiently to meet, the ecological and phenotypic similarity of these species may predict that the species merge into a single species. In contrast, classical hybrid speciation (right panel) produces one species of hybrid origin that may be ecologically distinct and reproductively isolated from both parental species and can hence coexist in sympatry with both. We exemplify hybrid speciation in Lake Victoria cichlids with *Mbipia mbipi* which may have resulted from hybridization between *M. lutea* and *P. pundamilia* (Keller, et al. 2013). Larger evolutionary distance between the parental species (illustrated by multiple intermittent speciation events) may facilitate the formation of a hybrid species as more incompatibilities are expected to segregate in the hybrid population, and the novel ecological and/or sexual

trait combinations may be more transgressive and differ from both parental species along more phenotypic axes of divergence (Selz, Lucek, et al. 2014; Selz, Thommen, et al. 2014). Hybrid speciation requires either an underutilized niche that the hybrid species can exploit better than either parental species or geographical distance to the parental species. In the case of "hybrid parallel speciation", on the other hand, two species are generated from a hybrid population that are reproductively isolated and ecologically distinct from each other but that are expected to be ecologically and phenotypically similar to the two parental species. Hybrid speciation and "hybrid parallel speciation" represent different outcomes of a continuum of possible ways by which admixture variation may be involved in speciation, varying in extent of ecological differentiation and reproductive isolation within the hybrid population and between the hybrid population and its parental species, mediated by the geographic context, niche availability, and the genetic similarity between the parental species.