Genomics of Parallel Ecological Speciation in Lake Victoria Cichlids

Joana Isabel Meier,[1,2,3] David Alexander Marques,[1,2,3] Catherine Elise Wagner,[†,1,3] Laurent Excoffier,[2,4] and Ole Seehausen*,[1,3]

[1]Aquatic Ecology and Evolution, Institute of Ecology and Evolution, University of Bern, Bern, Switzerland

[2]Computational and Molecular Population Genetics (CMPG), Institute of Ecology and Evolution, University of Bern, Bern, Switzerland

[3]Department of Fish Ecology and Evolution, Center for Ecology, Evolution and Biogeochemistry, EAWAG Swiss Federal Institute of Aquatic Science and Technology, Kastanienbaum, Switzerland

[4]Swiss Institute of Bioinformatics, Lausanne, Switzerland

[†]Present address: Biodiversity Institute & Botany Department, University of Wyoming, Laramie, WY, USA

*Corresponding author: E-mail: ole.seehausen@eawag.ch.

Associate editor: Deepa Agashe

Abstract

The genetic basis of parallel evolution of similar species is of great interest in evolutionary biology. In the adaptive radiation of Lake Victoria cichlid fishes, sister species with either blue or red-back male nuptial coloration have evolved repeatedly, often associated with shallower and deeper water, respectively. One such case is blue and red-backed *Pundamilia* species, for which we recently showed that a young species pair may have evolved through "hybrid parallel speciation". Coalescent simulations suggested that the older species *P. pundamilia* (blue) and *P. nyererei* (red-back) admixed in the Mwanza Gulf and that new "nyererei-like" and "pundamilia-like" species evolved from the admixed population. Here, we use genome scans to study the genomic architecture of differentiation, and assess the influence of hybridization on the evolution of the younger species pair. For each of the two species pairs, we find over 300 genomic regions, widespread across the genome, which are highly differentiated. A subset of the most strongly differentiated regions of the older pair are also differentiated in the younger pair. These shared differentiated regions often show parallel allele frequency differences, consistent with the hypothesis that admixture-derived alleles were targeted by divergent selection in the hybrid population. However, two-thirds of the genomic regions that are highly differentiated between the younger species are not highly differentiated between the older species, suggesting independent evolutionary responses to selection pressures. Our analyses reveal how divergent selection on admixture-derived genetic variation can facilitate new speciation events.

Key words: ecological speciation, hybridization, evolutionary genomics, cichlids.

Introduction

Understanding the genetic underpinnings of reproductive isolation and divergent adaptation is a major focus in speciation research. Some verbal models of speciation-with-gene-flow predict that genomic divergence in the face of gene flow is initially confined to very few small genomic regions under strong divergent selection (Wu 2001; Feder, Egan, et al. 2012; Via 2012). One idea is that diverging regions may eventually grow due to local reduction of gene flow, allowing nearby sites under weaker divergent selection to diverge too (Via and West 2008; Feder and Nosil 2010; Feder, Egan, et al. 2012; Via 2012). As the regions of divergence increase in number and size, genome-wide reduction of gene flow eventually results (Barton 1983; Wu 2001; Feder, Egan, et al. 2012; Feder, Gejji, et al. 2012; Flaxman et al. 2012, 2014). Taxa at early stages of speciation that still hybridize are particularly suitable study systems for speciation-with-gene-flow as they have not yet accumulated differences that arose after reproductive isolation is complete (Coyne and Orr 2004). Ideally, one would study replicate pairs of taxa that differ in nothing else but strength of reproductive isolation, in order to assess the factors influencing variation in the cessation of gene flow between emerging species (Martin et al. 2013; Riesch et al. 2017). Pairs of species that have evolved through parallel speciation offer such an opportunity, and have provided compelling evidence for a role of selection in speciation (Schluter and Nagel 1995; Rundle et al. 2000; Nosil et al. 2009; Johannesson et al. 2010).

In the young adaptive radiation of Lake Victoria cichlid fishes, replicates of similar species pairs are characterized by parallel divergence along the same major phenotypic and ecological axes (Seehausen 1996). Altogether the radiation comprises about 500 species (Seehausen 2002) that likely evolved in <15,000 years (Johnson et al. 2000; Stager and Johnson 2008). The evolution of hundreds of phenotypically and ecologically distinct species in rapid succession requires large amounts of genetic variation in traits relevant to ecological adaptation and reproductive isolation. In an earlier study (Meier, Marques, et al. 2017), we showed that an

ancient admixture event between two divergent cichlid lineages, providing such genetic variation, preceded the adaptive radiations of cichlids in Lake Victoria and other lakes in the region. During the process of adaptive radiation, species may fix different adaptive variants and thus lose genetic variation at sites conferring adaptation to different niches or reproductive isolation. Subsequent speciation events from any individual species would then become increasingly more difficult as the genetic variation gets depleted. Much of the ancestral variation may still be present in the adaptive radiation but now confined to individual species. Recurrent introgressive hybridization among some of the species may slow down the depletion of genetic variation at functionally relevant genes and thus facilitate further speciation events (Grant and Grant 1992; Seehausen 2004). Evidence for hybridization's contribution to speciation in Lake Victoria cichlids has been found for several species (Keller et al. 2013; Meier, Sousa, et al. 2017).

The Lake Victoria cichlid species pair *Pundamilia nyererei* and *P. pundamilia*, and other phenotypically similar species pairs, are particularly well-suited to study how selection and hybridization influence adaptation and speciation. These species coexist in full sympatry at different rocky islands in the open and offshore sectors of Lake Victoria, where they exhibit variable levels of genetic and phenotypic differentiation (fig. 1; Seehausen et al. 2008; Seehausen 2009). *P. nyererei* tends to live in deeper water than *P. pundamilia* where the light spectrum is more red-shifted, and it carries a red-shifted *LWS* opsin haplotype (Carleton et al. 2005; Seehausen et al. 2008), has red shifted color perception (Maan et al. 2006), males have red back nuptial coloration (fig. 1*b*), and females have a strong preference for red males (Maan et al. 2004; Seehausen 2009). *P. pundamilia* lives in shallower water, has blue male nuptial coloration (fig. 1*b*), a less red-shifted *LWS* opsin haplotype, blue-shifted color perception, and females prefer blue males (Seehausen and van Alphen 1998; Seehausen 2009). The two species also differ in morphology, for example, *P. pundamilia* has a longer head, wider and longer lower jaws, a longer snout, deeper cheeks, smaller eyes, and less teeth but more tooth rows in the upper oral jaw than *P. nyererei* (van Rijssel et al. 2018). These traits have a high heritability (Magalhaes et al. 2009).

The Mwanza Gulf, a fjord-like extension in the south of Lake Victoria, contains many islands that also harbor red-back and blue *Pundamilia* that occupy similar niches to *P. nyererei* and *P. pundamilia* in the main lake (fig. 1; Seehausen 1996; Seehausen et al. 2008). Females have strong mating preferences for males based on color (Seehausen and van Alphen 1998; Stelkens et al. 2008; Selz et al. 2014). The two species show morphological differences in the same direction as *P. nyererei* and *P. pundamilia* in head length, lower jaw length, snout length, and number of teeth in the upper oral jaws (van Rijssel et al. 2018). However, the red-backed (nyererei-like) and blue (pundamilia-like) populations in the central and north-western Mwanza Gulf differ in multiple details of coloration and morphology from *P. nyererei* and *P. pundamilia* (Seehausen 1996; Meier, Sousa, et al. 2017; van Rijssel et al. 2018). Some morphological traits differ between the species in opposite directions compared with *P. pundamilia* and

P. nyererei. For example, *P.* sp. "pundamilia-like" has larger eyes than *P.* sp. "nyererei-like", whereas *P. pundamilia* has smaller eyes than *P. nyererei* (van Rijssel et al. 2018). Based on microsatellite data, the two Mwanza Gulf species are genetically more similar to each other than either of them is to *P. nyererei* and *P. pundamilia* (Seehausen et al. 2008). Females of *P.* sp. "nyererei-like" from Python Island in the Mwanza Gulf prefer conspecific males over *P. nyererei* males from outside of the Mwanza Gulf (Selz et al. 2016), suggesting behavioral reproductive isolation between the central and north-western Mwanza Gulf species and the species in the main lake (fig. 1).

We recently investigated the evolutionary history of these species with demographic modeling of *P. pundamilia* and *P. nyererei* from Makobe Island in the open lake north of the Mwanza Gulf and *P.* sp. "pundamilia-like" and *P.* sp. "nyererei-like" from Python Island in the central Mwanza Gulf (fig. 1; Meier, Sousa, et al. 2017). We found that *P.* sp. "pundamilia-like" and *P.* sp. "nyererei-like" are of hybrid origin between *P. pundamilia* and *P. nyererei* (Meier, Sousa, et al. 2017). Under the best-supported model, *P. pundamilia* established first in the Mwanza Gulf, and later *P. nyererei* individuals colonized the Mwanza Gulf and admixed with the local *P. pundamilia* population. Shortly thereafter, the hybrid population speciated into a shallow living species with blue males (*P.* sp. "pundamilia-like") and a deeper living species with red-backed males (*P.* sp. "nyererei-like"). Some of the allelic variation that facilitated the evolution of the new red-backed species, which colonized the deep-water niche, may have arrived from *P. nyererei*. In this study, we explore the sorting of admixture-derived alleles into the new species and ask to what extent phenotypic parallelism between the older and the younger species pair is reflected in genome-wide patterns of species divergence. The younger species in the Mwanza Gulf may have emerged through "respeciation" (sensu Gilman and Behm 2011), where the new species diverge in all or a subset of the key genomic regions that the older species pair had already experienced divergence in. Alternatively, the evolution of these species may resemble a scenario of "hybrid parallel speciation", where the new species parallel the older ones phenotypically and in some of the divergent genomic regions, but additionally diverge in many other regions in the genome that are not divergent in the older species pair.

Candidate genes that may have facilitated *Pundamilia* speciation include the red-sensitive (*LWS*) opsin gene, which is involved in adaptation to deep versus shallow water light environments (Seehausen et al. 2008), the blue-sensitive opsin gene *SWS-2A* (Seehausen et al. 2008), genes mediating red versus blue male nuptial coloration (Magalhaes and Seehausen 2010), genes contributing to morphological (Magalhaes et al. 2009; van Rijssel et al. 2018) or metabolic differences involved in diet shifts (Seehausen et al. 1998), or genes conferring female preference for red or blue male nuptial coloration (Svensson et al. 2017). The hybrid parallel speciation scenario, where the second speciation event is facilitated by introgression of *P. nyererei* genes into a local population of *P. pundamilia* (Meier, Sousa, et al. 2017),
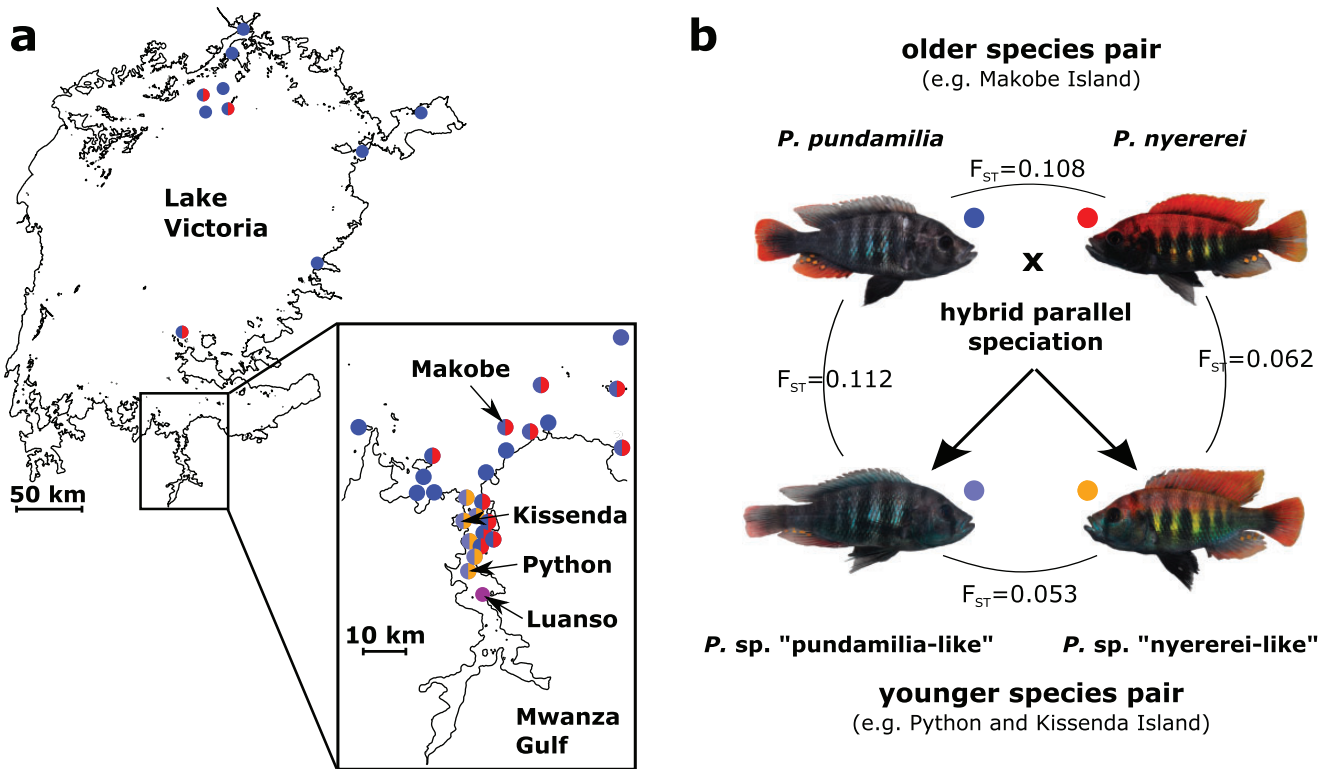
Fig. 1. Distribution and sampling sites of red and blue *Pundamilia* cichlids. (*a*) Map of Lake Victoria with known records (Meier, Sousa et al. 2017) of the older species pair, *P. pundamilia* (blue) and *P. nyererei* (red), and of the younger species pair, *P.* sp. "pundamilia-like" (light blue) and *P.* sp. "nyererei-like" (orange). Sites harboring two species are shown as two-colored dots. The inset shows the Mwanza Gulf with the study sites, including Luanso Island (purple), which harbors a color-polymorphic *Pundamilia* population that is not differentiated. (*b*) Photos of representative males of each species at the two main study sites, Makobe and Python Island, and mean F_{ST} estimates between the species. The younger species pair likely evolved from a hybrid swarm of the two older species through "hybrid parallel speciation" (Meier, Sousa et al. 2017).

predicts that genomic regions carrying such genes should be differentiated in the original species pair and be among the most strongly differentiated genes in the younger pair. However, there should also be many other differentiated regions between the younger species that are not shared with the original older species pair. These other regions would have become differentiated during the new speciation process. In contrast, "respeciation" in the sense of restoration of the original species differences after massive hybridization would imply that regions of strong genomic differentiation in the young species pair are a subset of the genomic regions of high differentiation in the original species pair.

Here, we use whole-genome resequencing data from the sympatric *Pundamilia* species pairs at Makobe Island (the older species pair) and at Python Island (the very recent species pair) (Meier, Sousa, et al. 2017) to characterize the genomic patterns and potential parallelism of divergence in the species pairs, assess the role of admixture-derived variation, and distinguish between the respeciation and the hybrid parallel speciation scenarios. We find that many of the genomic regions that are most strongly differentiated between the older *P. nyererei* and *P. pundamilia* are also differentiated between the younger *P.* sp. "nyererei-like" and *P.* sp. "pundamilia-like" and mostly show high allele sharing between older and younger species of the same color type. However, two-thirds of the high differentiation regions in

the young species pair constitute novel candidate targets of divergent selection in this pair, as they show no signature of selection in the original species pair, consistent with hybrid parallel speciation.

Results

Nucleotide Diversity and Divergence

We used whole-genome resequencing data of four individuals each of *P. nyererei* and *P. pundamilia* from Makobe Island (the older original species) and of *P.* sp. "nyererei-like" and *P.* sp. "pundamilia-like" from Python Island (the younger species in the Mwanza Gulf). By mapping the reads to the anchored *P. nyererei* reference genome (Feulner et al. 2018), we obtained 630 million base pairs (Mb) including 5,199,846 biallelic SNPs sequenced in at least half of the individuals with a minimum of ten reads each (mean depth of coverage is 16–29 reads, supplementary table S1, Supplementary Material online). The mean missing data proportion per site was 3.1%. The mean proportion of biallelic SNPs among all sites in the data set combining all populations was 0.44% (excluding sites with missing data, supplementary table S1, Supplementary Material online) and not much lower in single species data sets (0.36–0.41%, supplementary table S1, Supplementary Material online). The mean weighted F_{ST} was 0.108 between *P. pundamilia* and *P. nyererei* at Makobe, 0.053 between *P.* sp.

"pundamilia-like" and *P.* sp. "nyererei-like" at Python (fig. 1) based on sites without missing data (on average 2.1 million SNPs). The vast majority of SNPs were shared across multiple or all species. Of the SNPs with maximum one individual missing per species and a minor allele count of three across all four species, only 1% of the variants were unique to *P. pundamilia* Makobe, 0.31% to *P. nyererei* Makobe, 0.36% to *P.* sp. "pundamilia-like" Python, and 0.24% to *P.* sp. "nyererei-like" at Python Island. 4.2% of these variants were only found in the species at Makobe and 4.4% only in the species at Python Island.

One-Third of the High Differentiation Regions Are Shared between Species Pairs

To assess if the genomic differentiation between *P.* sp. "nyererei-like" and *P.* sp. "pundamilia-like" paralleled that between *P. nyererei* and *P. pundamilia*, we identified highly differentiated genomic regions for Python and Makobe Island separately, using a Hidden Markov Model (HMM, Baum and Petrie 1966) approach. For each species pair, we computed a null model expectation for genome-wide differentiation (F_{ST}) from simulations under the best demographic model in Meier, Sousa, et al. (2017) (supplementary fig. S1, Supplementary Material online). Based on these simulated null distributions, we computed z-scores (normalized P values) for our observed F_{ST} values averaged across 15 SNPs. Given that the parameters of the demographic model were estimated using a data set that contained both neutrally and nonneutrally evolving genomic regions (Meier, Sousa, et al. 2017) and thus outlier detection using this null model is highly conservative. An HMM with two hidden states corresponding to "normal" and "high differentiation" was run over the z-scores. Of the 137,550 15-SNP windows in the Makobe data set and 112,478 15-SNP windows in the Python data set, 5.4% and 5.9%, respectively, had a z-score >1.65 (corresponding to a one-sided P value of 0.05), and 7.8% and 7.1% were assigned to the high differentiation state in the Makobe and Python data sets, respectively (table 1).

The mean F_{ST} of windows assigned to the highly differentiated state was 4.5 and 6.7 times as high as the mean of the other windows at Makobe Island (0.296 vs. 0.066) and Python Island (0.207 vs. 0.031), respectively (table 1 and supplementary fig. S1, Supplementary Material online). When highly differentiated regions were excluded, the genome-wide mean F_{ST} from SNPs pruned for high linkage ($r^2 > 0.5$) was 0.031 with a 99% bootstrap confidence interval (CI) of 0.030–0.032 for the young species pair at Python Island and 0.058 (99% CI = 0.056–0.059) for the older species pair at Makobe Island, closely resembling previously published F_{ST} estimates from microsatellites (Seehausen et al. 2008). This suggests that both pairs of sympatric species are experiencing some genome-wide reproductive isolation also outside the high differentiation regions.

We combined adjacent high differentiation windows to highly differentiated genomic regions (HDRs). This approach resulted in 346 Makobe HDRs and 365 Python HDRs, respectively (table 1). Makobe HDRs were larger than Python HDRs (on an average 268 vs. 196 kb, two-sided Mann–Whitney U-test, P value = 0.02, table 1 and supplementary fig. S2,

Supplementary Material online). By extracting the parts of the HDRs that overlap between Makobe and Python Island, we obtained 120 shared HDRs (SHDRs) that were widely scattered across the genome (fig. 2, blue vertical lines). Of the 346 Makobe HDRs, 103 overlapped with at least one Python HDR (29.8%), indicating that only a subset of the highly differentiated regions of the original species pair is also differentiated between the younger species at Python Island. Of the Python HDRs, a very similar fraction of 28.8% (105/365 HDRs) overlapped with at least one Makobe HDR, indicating that many unique genomic regions show signatures of divergent selection in the younger species pair that were not divergent between the older species. The proportion of HDRs shared between both species pairs is robust to different approaches and window sizes used to identify HDRs (supplementary table S2, Supplementary Material online). All shared HDRs together covered 21.1 Mb, which is 22.6% and 29.3% of the HDR lengths at Makobe and Python Islands, respectively. Both the length distribution of HDRs (supplementary fig. S2a and b, Supplementary Material online) and that of SHDRs were strongly right-tailed (supplementary fig. S2c, Supplementary Material online), the latter with mean 176 kb, median 91 kb, and a maximum length of 1.5 Mb. The overlap of HDRs from Python and Makobe is much larger than expected in a block-permuted data set (supplementary fig. S3, Supplementary Material online).

We computed F_{ST} estimates for both species pairs in non-overlapping 20-kb windows to compare F_{ST} values of the same genomic regions between the two pairs. The HDRs unique to one species pair did not show higher F_{ST} values than the genome-wide background in the other species pair (supplementary fig. S4, Supplementary Material online). The F_{ST} estimates for both species pairs were slightly and significantly higher in shared HDRs than in HDRs unique to either of the species pairs (supplementary fig. S4, Supplementary Material online), suggesting that many of the most strongly differentiated regions between *P. nyererei* and *P. pundamilia* were also targeted by divergent selection in the younger species pair at Python Island.

Highly Differentiated Genomic Regions Show Signatures of Selection in Multiple Statistics

We further characterized highly differentiated genomic regions by computing additional statistics for selection, nucleotide diversity, recombination rate, and introgression in windows of 20 kb. We obtained 32,864 20-kb windows with an average of 17,508 covered sites (median: 18,286) and 130 SNPs. For Makobe and Python Island separately, we compared 20-kb windows overlapping with HDRs (HDR windows) with windows overlapping with shared HDRs (shared HDR windows), and with windows not overlapping with any HDRs representing the genome-wide background (nonHDR windows).

We tested whether HDRs were enriched for signatures of selection using two distinct "extended haplotype length" (EHL) tests. Both tests are based on the prediction that in a selective sweep, a variant rises to high frequency so rapidly that linkage disequilibrium with neighboring polymorphisms is not disrupted by recombination, giving rise to haplotypes of

Table 1. Properties of 15-SNP windows assigned to the high differentiation state (HD) and of genomic regions of high differentiation (HDR) between *Pundamilia nyererei* and *P. pundamilia* at Makobe (older species pair) and *P.* sp. "nyererei-like" and *P.* sp. "pundamilia-like" at Python Island (younger species pair), as identified with the HMM.

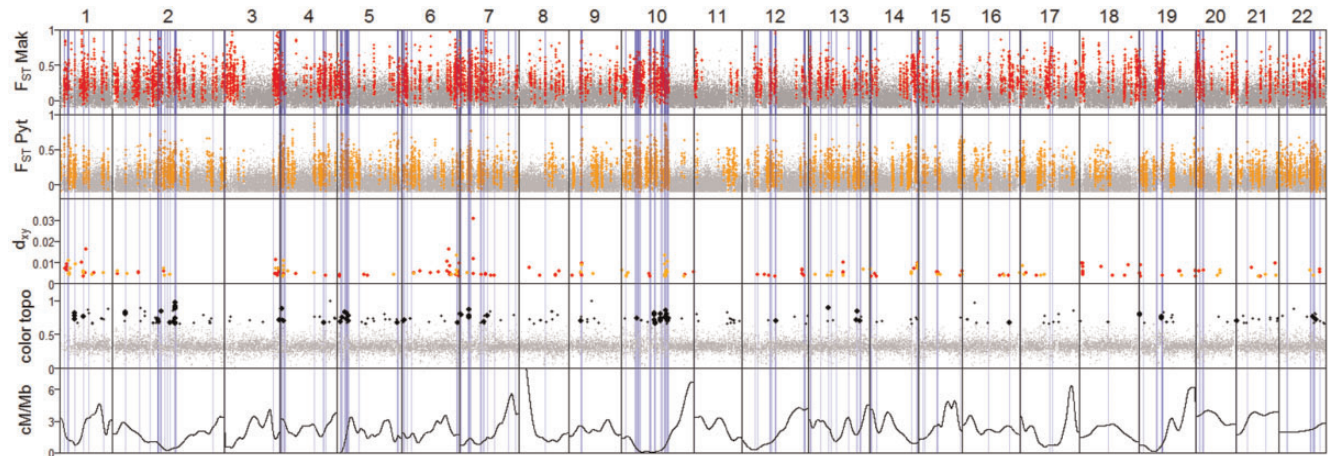|	Number of HD Windows/Total	Mean F_{ST} in HD/Other Windows	Count of HDRs	Mean Length of HDRs (kb)	Mean Window Counts per HDR	Cumulative Length of HDRs (Mb)	HDR Overlapping with At Least One HDR in the Other Species Pair
Makobe	8,709/111,901 (7.8%)	0.296/0.066	346	268	25	93.0	103 (29.8%)
Python	9,811/136,961 (7.2%)	0.207/0.031	365	196	27	71.8	105 (28.8%)

Fig. 2. Genomic regions that are highly differentiated in the younger and the older species pair are numerous and widespread across the genome. F_{ST} averages of 15-SNP windows between the older species pair (Makobe, top row) and the younger species pair (Python, second row) are shown along the chromosomes (linkage group numbers from Feulner et al. 2018), whereby windows assigned to the "high differentiation state" are shown in red (Makobe) or orange (Python). Light blue vertical bars indicate the locations of shared HDRs. Other rows show d_{XY} outliers for Makobe (red) and Python (orange), weights of the "color topology", indicating how strongly the species cluster by male nuptial coloration, and predicted recombination rates in centimorgans per megabase (cM/Mb). Color topology windows (color topology weight > 66%) indicating strong genomic parallelism between the species pairs are shown in black. The 71 color topology windows highlighted with larger black symbols coincide with shared HDRs and are thus considered candidates of parallel divergent selection on shared genetic variation.

extended length. The first of these tests that we computed is the Cross Population Extended Haplotype Homozygosity statistic (XP-EHH, Sabeti et al. 2007) which compares the haplotype lengths between the sympatric sister species. A selective sweep in one of the species would lead to longer haplotypes compared with the other species (Sabeti et al. 2007). In addition, we computed the integrated haplotype score which compares the lengths of different haplotypes within a population (iHS, Voight et al. 2006). This test has most power during an ongoing sweep, when the haplotype under positive selection is expected to be much longer than the other haplotypes but is not fixed yet. We used absolute iHS as we did not have information on the derived or ancestral state of the alleles, and absolute XP-EHH as we did not distinguish in which of the diverging species in a pair selection acted. We found that compared with regions that are not highly differentiated (nonHDRs), HDRs unique to the Makobe species pair had generally higher XP-EHH and iHS values computed for the Makobe species but not for the Python species. Similarly, compared with nonHDRs, HDRs unique to the Python species pair had on an average higher XP-EHH and iHS values computed for the Python species but not for the Makobe species (supplementary fig. S5, Supplementary Material online).

Of the genomic regions that were highly differentiated either only in the older or only in the younger species pair, 61% and 49%, respectively, showed significant extended haplotype length statistics (significant iHS in at least one species and/or significant XP-EHH) in at least 20% of the spanned windows (fig. 4c and supplementary fig. S6, Supplementary Material online, excluding HDRs without EHL information). The fraction of HDRs that are EHL outlier HDRs is 3.0 (Makobe) or 2.5 (Python) times as high as the average fraction obtained from block-permuted null distributions (supplementary fig. S6, Supplementary Material online). Shared HDRs were even more frequently outliers for EHL statistics (3.3× and 2.6× enrichment for Makobe and Python EHL signatures, respectively, supplementary fig. S6, Supplementary Material online). However, HDRs unique to the Makobe species were not enriched for Python EHL outliers, and HDRs unique to the Python species pair were not enriched for Makobe EHL outliers (supplementary fig. S6, Supplementary Material online).

As a further test of divergent selection, we computed the difference in nucleotide diversity (Δpi) between the species in windows of 20 kb. A high absolute $\Delta\pi$ indicates that a selective sweep reduced the nucleotide diversity in one population. We considered windows with a $\Delta\pi$ value above the 90th

quantile as $\Delta\pi$ outlier windows providing additional evidence of divergent selection. Makobe HDRs were enriched for Makobe $\Delta\pi$ outlier windows and Python HDRs were enriched for Python $\Delta\pi$ outlier windows supporting that HDRs were formed by divergent selection (fig. 4c and supplementary fig. S6, Supplementary Material online). In contrast, HDRs unique to one species pair did not show enrichment for high $\Delta\pi$ in the other species pair (fig. 4c and supplementary fig. S6, Supplementary Material online), supporting absence of divergent selection in the other species pair.

As yet a further test, we explored signatures of selection using a measure of absolute sequence divergence. In contrast to relative measures of divergence such as F_{ST}, which compare genetic differentiation between populations to genetic variation within populations, absolute measures of divergence are less affected by forces decreasing within-population variation such as background selection (Nachman and Payseur 2012; Cruickshank and Hahn 2014). Nei's measure of absolute divergence (d_{XY}) calculated in windows of 20 kb was highly correlated with within-species diversity (supplementary fig. S7, Supplementary Material online, Pearson's correlation [r^2] between d_{xy} and the average of the two π estimates is 0.81 for the older species pair at Makobe and 0.92 for the younger species pair at Python). This is in line with the prediction that for comparisons of very recently diverged taxa, d_{XY} mostly correlates with mutation rates and levels of ancestral variation (Nachman and Payseur 2012; Cruickshank and Hahn 2014; Riesch et al. 2017). We identified outlier windows where high d_{xy} values are unlikely to be due to increased mutation rate: windows with both d_{XY} and d_{XY}/π values >1.5 times the interquartile range (IQ) above the third quartile (Q3). Using this two-pronged approach, we identified 134 and 83 d_{XY} outlier windows for the older and the younger species pair, respectively (fig. 2). Of these windows with high d_{XY} between the older species, 83% were distributed across 63 different HDRs, of which 17 were shared HDRs. Similarly, 84% of the windows with high d_{XY} in the younger species pair were located in 46 HDRs, of which 14 were shared with the older species pair. HDRs unique to either one of the species pairs were strongly enriched for outlier windows of d_{XY} in the respective species pair compared with block-permuted null expectations (4× for the older species pair, 6× for the younger species pair, fig. 4 and supplementary fig. S6, Supplementary Material online). However, HDRs unique to one species pair did not show enrichment for d_{xy} outlier windows in the other species pair (fig. 4), confirming that these regions were only under divergent selection in the focal species pair. Shared HDRs were most strongly enriched for d_{XY} outliers with 6× and 9× enrichment compared with the mean of block-permuted shared HDRs for Makobe and Python d_{XY} outliers, respectively (fig. 4 and supplementary fig. S6, Supplementary Material online).

Finally, as yet one more signature of selection that also captures regions with ongoing and completed selective sweeps, we computed Tajima's D in 20-kb windows. After a sweep, the beneficial haplotype will reach high prevalence and new mutations will only arise slowly and initially be at low frequency. This excess of rare alleles renders Tajima's D values negative in sweep regions. However, we caution that in the very young species pair at Python Island, new mutations that arose after the onset of speciation are expected to be very few if any and Tajima's D may not convey much of a selection signal. If at least 20% of the windows spanned by the HDR had strongly negative Tajima's D values below the fifth quantile for one or both species, we considered the HDR to be a Tajima's D outlier HDR (fig. 4c and supplementary fig. S6, Supplementary Material online). As expected, we found a strong enrichment of Tajima's D outlier HDRs in the older original species pair (3×) but only a weak enrichment in the young species pair (fig. 4c and supplementary fig. S6, Supplementary Material online). HDRs unique to the original species pair were also slightly enriched for strongly negative Tajima's D in the younger species pair, which is likely a legacy of past selection in the original species pair (fig. 4c and supplementary fig. S6, Supplementary Material online). This might generally be a problem in cases of recent divergence events from populations of admixed ancestry and warrants caution in the interpretation.

Of the HDRs unique to the older or to the younger species pair with information for all five statistics of selection (XP-EHH, iHS, d_{XY}, $\Delta\pi$, Tajima's D) in addition to F_{ST}, 72% or 64%, respectively, were significant for at least one other selection statistic. The percentage of HDRs unique to one species pair showing additional evidence for selection in that species pair was much higher than expected from block-permuted null distributions (fig. 4a and b; supplementary fig. S6, Supplementary Material online). The shared HDRs were even more strongly enriched for additional evidence of selection (fig. 4a and b; supplementary fig. S6, Supplementary Material online). Of the 120 shared HDRs, 100 showed additional significant selection statistics in the original species pair at Makobe Island and 95 in the young species pair at Python Island (113 in at least one of the species pairs). Most HDRs showed additional signatures of selection in multiple statistics (supplementary fig. 8, Supplementary Material online). The different selection statistics thus strongly support the action of divergent selection in driving the high differentiation in HDRs. In contrast, HDRs unique to one species pair are not enriched for significant selection statistics in the other species pair, as shown by block-permutation (fig. 4a and b). This indicates that HDRs unique to one species pair are indeed not under divergent selection in the other species pair.

Genomic Regions under Divergent Selection Are Associated with Low Recombination Rates

We tested if highly differentiated genomic regions were associated with lower than expected recombination rates. Regions under strong divergent selection may be associated with low recombination where selection may be more efficient due to reduced effective gene flow and recombination between co-adapted alleles (Rieseberg 2001; Butlin 2005; Hoffmann and Rieseberg 2008; Bürger and Akerman 2011; Yeaman and Whitlock 2011; Nachman and Payseur 2012). However, similar patterns of strong differentiation in low recombination regions can be caused by purifying selection against recurrent deleterious mutations (termed background selection,

Charlesworth et al. 1993; Nachman and Payseur 2012) or against maladaptive introgressed alleles (Martin and Jiggins 2017). In the absence of gene flow, background selection is likely to be the major driver of the commonly observed association between high differentiation and low recombination rates (Nachman and Payseur 2012; Renaut et al. 2013; Cruickshank and Hahn 2014; Burri et al. 2015). However, in cases of recent and sympatric divergence with ongoing gene flow, such as the young *Pundamilia* species pair at Python Island, background selection is unlikely to cause regions of low recombination to diverge between the species (see also Cruickshank and Hahn 2014; Burri 2017).

We found that the average recombination rates were lower in HDRs compared with the rest of the genome (fig. 3). Shared HDRs were even more strongly associated with low recombination rates (fig. 3). This finding was consistent if we restricted the analyses to HDRs with additional evidence for divergent selection (supplementary fig. S9, Supplementary Material online) suggesting that divergent selection rather than background selection is driving the association. We found a genome-wide negative association between recombination rate and multiple statistics of selection (iHS, XP-EHH, Δpi), which are less likely to be affected by background selection than F_{ST} (supplementary fig. S10, Supplementary Material online). Outlier windows of high absolute divergence (d_{xy}) are unlikely to be caused by background selection and also had a reduced mean recombination rate compared with the rest of the windows (1.99 vs. 2.32 cM/Mb for the older species pair and 1.91 vs. 2.32 cM/Mb in the younger pair).

In addition, if background selection was driving the similarity in genomic differentiation landscapes between species pairs (e.g. Renaut et al. 2013; Burri et al. 2015) and contributing to shared HDRs in low recombination regions, we would expect high correlation of F_{ST} window estimates between the species pairs. However, we found low overall correlation of genetic differentiation patterns between the species pairs even though the recombination map is likely highly similar (supplementary fig. S11, Supplementary Material online). Using the RAD data from Meier, Sousa, et al. (2017), we computed F_{ST} estimates between the older species at Makobe Island, the younger species at Python Island and at another island (Kissenda Island, fig. 1), and between each species at Python Island and an allopatric *Pundamilia* population at a fourth island (Luanso Island in the South of the Mwanza Gulf, fig. 1). We tested if 20-kb windows that are highly differentiated between the older species pair at Makobe Island (mean $F_{ST} > 0.2$) are also more strongly differentiated between allopatric species that are unlikely to experience parallel selection pressures. We found that in these windows, differentiation of either species from Python Island against the allopatric Luanso population was not higher than expected from block-permutation null distributions (supplementary fig. S10c, Supplementary Material online). In contrast, windows that were highly differentiated between the older species at Makobe Island also showed elevated differentiation between the sympatric younger species at both islands with sympatric species pairs
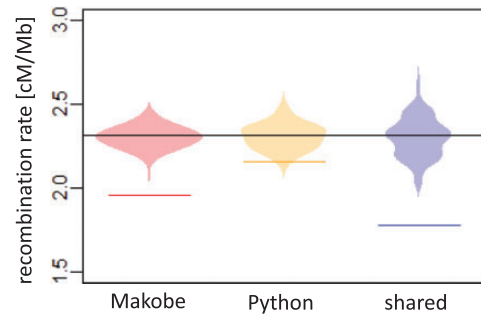

Fig. 3. Highly differentiated genomic regions are associated with low recombination rates. The mean recombination rate in highly differentiated regions (HDRs) unique to the Makobe species pair (red horizontal line), unique to the Python species pair (orange line), or shared by both (blue line) are significantly lower than block-permuted null distributions (bean plots, *P* values: Makobe HDR: <0.01, Python HDR: 0.02, shared HDR: <0.01). The black horizontal line indicates the overall mean recombination rate.

(Python and Kissenda Island, supplementary fig. S11c, Supplementary Material online, permutation test: $P < 0.001$). This provides additional evidence that parallel divergent selection and not background selection drives the elevated differentiation in the sympatric species pairs.

Parallel Selection on Same Alleles in the Old and Young Species Pairs

To test if species with the same male nuptial color syndrome (red-back vs. blue) share more alleles in HDRs than species with different male color syndromes, we computed f_d (Martin et al. 2015), a measure of excess allele sharing in windows of 20 kb. We found that f_d among same-color species is much higher in shared HDR windows than in other windows, indicating high allele sharing between same-color species. Mean $f_d$nyer, measuring excess allele sharing between *P. nyererei* with *P.* sp. "nyererei-like" relative to *P.* sp. "pundamilia-like", was 0.27 in shared HDRs, but only 0.12 in the rest of the genome. Excess allele sharing between *P. pundamilia* and *P.* sp. "pundamilia-like" compared with *P.* sp. "nyererei-like" ($f_d$pund) was on an average 0.18 in shared HDRs but only 0.08 outside of shared HDRs. Shared HDRs, were strongly enriched for top 10% $f_d$nyer and top 10% $f_d$pund windows (fig. 4d and e), which indicates that in shared HDRs divergent selection in the younger species pair had acted on admixture-derived alleles. Of the shared HDRs with available estimates, ~60% overlapped with at least one top 10% $f_d$nyer or $f_d$pund window (fig. 4d and e).

As another way to identify genomic windows with parallel allele frequency differences between the sympatric species at Makobe and Python Island, we assessed the most likely species tree topology in nonoverlapping 20-kb windows using TWISST (Martin and Van Belleghem 2017). Overall, the "geography topology" which groups the species by island (Python or Makobe Island, respectively) is best supported (supplementary fig. S12, Supplementary Material online). However, in shared HDRs, the "color topology", where species are grouped by male nuptial color syndrome ([*P. nyererei* + *P.* sp. "nyererei-like"],[*P. pundamilia* + *P.* sp. "pundamilia-like"])

c

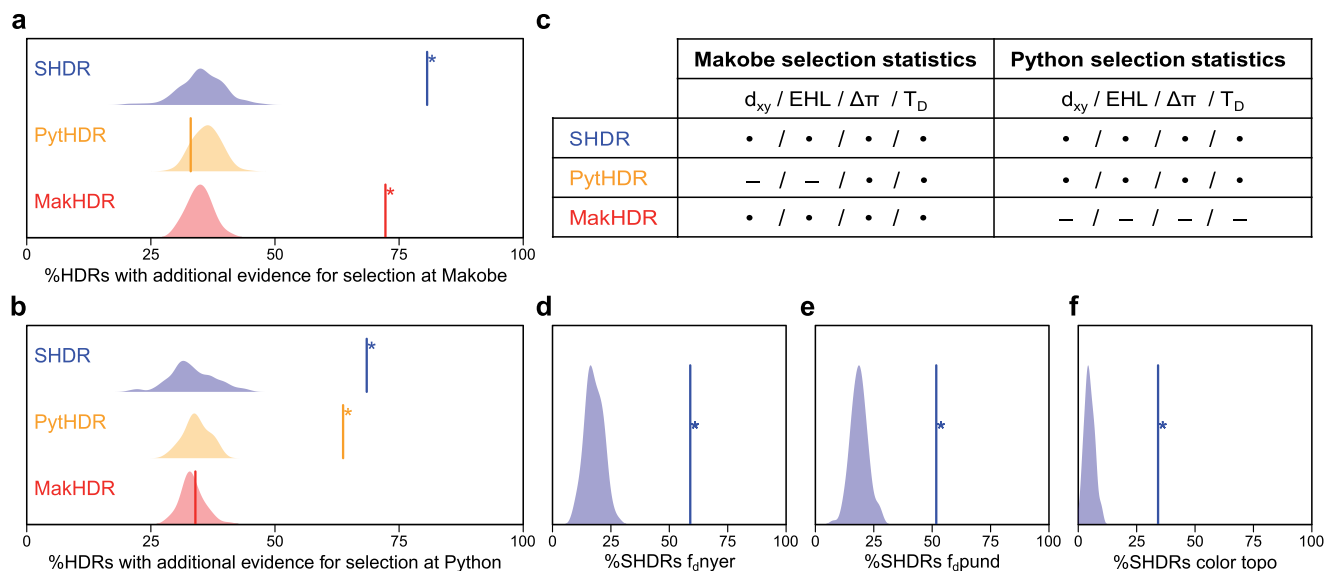|	Makobe selection statistics	Python selection statistics
	d_{xy} / EHL / $\Delta\pi$ / T_D	d_{xy} / EHL / $\Delta\pi$ / T_D
SHDR	• / • / • / •	• / • / • / •
PytHDR	– / – / • / •	• / • / • / •
MakHDR	• / • / • / •	– / – / – / –

Fig. 4. Enrichment of HDRs with evidence for selection. In each panel except for (*c*), we show the percentages of HDRs with significant statistics (vertical lines) compared with block-permuted null distributions (density plots). In (*a*) and (*b*), we show the proportion of HDRs with additional evidence for selection in the Makobe (*a*) or Python (*b*) species pair among HDRs unique to the older species pair at Makobe Island (MakHDR, red), among HDRs unique to the younger species pair at Python Island (PytHDR, orange), and among shared HDRs (SHDR, blue). (*a*) At Makobe Island, the proportion of MakHDRs and SHDRs with additional evidence for selection differs strongly from the null distributions, confirming the action of divergent selection generating elevated differentiation in these regions. In contrast, the proportion of PytHDRs with signatures of selection in the Makobe species pair is not different from null expectation, confirming the absence of selection in the Makobe species pair at genomic regions that are highly differentiated only in the Python Island species pair. (*b*) At Python Island, we see a parallel pattern to that in (*a*): PytHDRs and SHDRs are enriched for selection statistics in the Python species pair, but no such enrichment is found in the MakHDRs, providing evidence that selection in HDRs unique to one species pair is restricted to that species pair. (*c*) Most selection statistics of a given species pair only show enrichment in the HDRs of that species pair. We here show significant enrichment (dot) if the proportion of HDRs spanning an outlier window is higher than the null distribution; the minus symbol indicates nonsignificant enrichment (see supplementary fig. S6, Supplementary Material online, for more details). Selection statistics include d_{XY}, extended haplotype length tests (EHL, including iHS for both species and XP-EHH), difference in nucleotide diversity between the species ($\Delta\pi$), and low Tajima's D (T_D) in one or both species. (*d–f*) Shared HDRs show evidence for parallel selection leading to sorting of admixture-derived allelic variation in the younger species pair. This is shown by the highly significant proportion of SHDRs that show excess allele sharing between *P. nyererei* and *P.* sp. "nyererei-like" (d, $f_d$nyer) or between *P. pundamilia* and *P.* sp. "pundamilia-like" (e, $f_d$pund). Similarly, the proportion of SHDRs containing at least one "color topology window", whereby species of the same male nuptial coloration cluster together, strongly exceeds null expectations (f, see also, supplementary fig. S12, Supplementary Material online).

is more frequent than the "geography topology", consistent with parallel selection on shared genetic variation (supplementary fig. S12, Supplementary Material online). Shared HDRs had higher weights for the color topology and lower weights for the geography and the random topology compared with the genome-wide average (permutation test, $P < 0.001$ for all three topologies). We considered windows where the "color topology" weight exceeds 66% (twice as high as both other possible topologies together) as "color topology windows". As the sharing of variation is likely due to the admixed ancestry of the younger species, "color topology windows" indicate regions where the admixture variation was sorted between the species in the younger species pair under selection pressures paralleling those experienced by the species in the older species pair. The proportion of shared HDRs containing such a "color topology window" was much higher than expected from block-permuted null distributions (34%, $P < 0.01$, fig. 4f).

Shared HDRs with at least one "color topology window" were considered candidate regions of parallel selection on admixture-derived genetic variation (fig. 4f). We identified 41 such candidate regions containing 71 "color topology windows" (fig. 2). These regions overlapped with 182 known genes (supplementary table S3, Supplementary Material online, see also a visualization of the SHDRs with gene info on Dryad doi:10.5061/dryad.8p6m07g). These genes were spread across 25 of the 41 candidate regions of parallel selection. Among these candidate genes, 27 genes are involved in visual perception or nervous system development and may thus contribute to divergent adaptation to different light regimes at different water depths and to behavioral reproductive isolation based on male nuptial coloration. Four genes are associated with pigmentation and might confer differences in male nuptial coloration. Three genes are opsin genes known to be under divergent selection between the species and are described in more detail in the next section. Twenty-three other candidate genes are associated with carbohydrate, lipid, and folic acid metabolism. They may be involved in diet shifts between the *Pundamilia* species (Bouton et al. 1997). Fifteen further candidate genes are involved in immunity and may thus confer adaptation to divergent parasite pressures at different water depths (Maan et al. 2008). Eight candidate genes are associated with limb, bone and cartilage development and may contribute to morphological differences between the
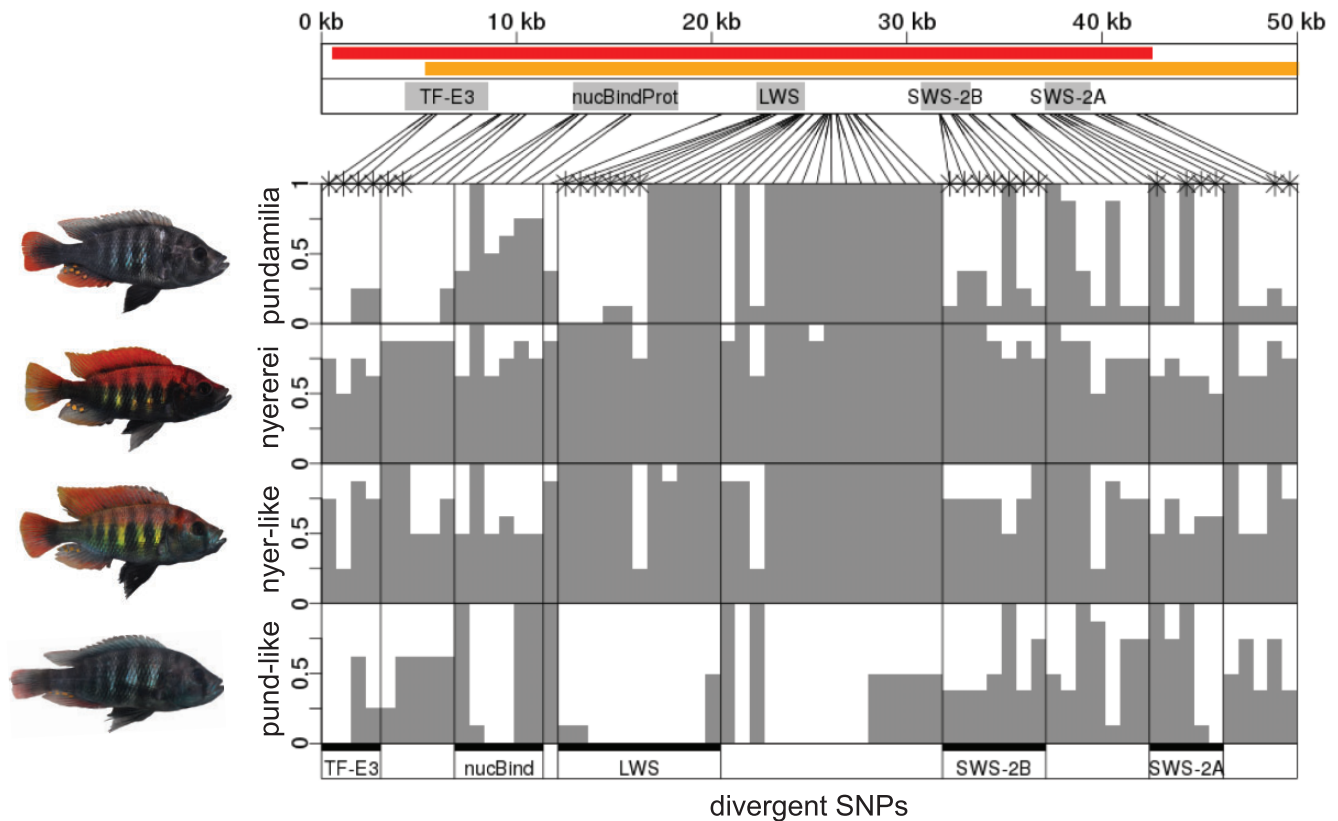
8

FIG. 5. Allele frequency differences in the genomic region of the long wavelength-sensitive (*LWS*) opsin gene. The region around the *LWS* opsin gene is located on linkage group 13 (Feulner et al. 2018) starting at position 26,306,571, corresponding to the beginning of scaffold 177 of the reference assembly of Brawand et al. (2014). The red and orange horizontal bars show the positions of the highly differentiated regions at Makobe and Python Island, respectively, whereas the light gray rectangles indicate gene positions. The first gene codes for the E3-like transcription factor (*TF-E3*), the second gene for a guanine nucleotide binding protein (nucBindProt), and the three remaining genes are cone opsin genes for red vision (*LWS*), blue vision (*SWS-2A*) and violet vision (*SWS-2B*). The lower panel shows allele frequencies in *P. pundamilia* (pundamilia) and *P. nyererei* (nyererei) at Makobe Island, and *P.* sp. "nyererei-like" (nyer-like) and *P.* sp. "pundamilia-like" (pund-like) at Python Island. Each vertical bar represents an individual SNP with an allele frequency difference between the species at Makobe and or Python Island of at least 0.4. The frequency of the major allele in *P. nyererei* is shown (in gray) for each species ranging from 0 (allele putatively absent) to 1 (allele putatively fixed). Sites with parallel allele frequency differences between the species at Makobe and Python Island are highlighted with stars on top. Note that 21 of 25 parallel sites are located in genes, whereas intergenic regions contain mostly sites with allele frequency shifts that are not parallel between the older and the younger species pair.

species (Seehausen et al. 1998; Magalhaes et al. 2009; van Rijssel et al. 2018). Fifteen other genes are associated with divergence in gene expression (transcription factors). Functional validation in further studies is required for most of these candidate genes of parallel divergence, including 36 genes currently lacking functional annotation, three candidate genes with a wide array of functions (supplementary table S3, Supplementary Material online), and the 16 candidate regions currently lacking known genes.

Divergence in the *LWS* Opsin Gene Region

The long wavelength sensitive (*LWS*) opsin gene is known to be importantly involved in adaptation to the different ambient light spectra at different water depths and is also correlated with male nuptial coloration and with female mate choice based on male nuptial coloration in *Pundamilia* cichlids (Carleton et al. 2005; Maan et al. 2006; Seehausen et al. 2008), and has recently been shown to directly determine variation in mating preferences in medaka (Kamijo et al.

2018). Seehausen *et al.* (2008) showed that *P. pundamilia* and *P.* sp. "pundamilia-like" predominantly have the P haplotype, whereas *P. nyererei* and *P.* sp. "nyererei-like" are almost fixed for the H haplotype, whose protein light absorption spectrum is red-shifted by 15 nm relative to the P haplotype. This led Meier, Sousa, et al. (2017) to hypothesize that the *LWS* opsin gene may represent a candidate gene for introgression from *P. nyererei* into Python *Pundamilia* that would have facilitated the evolution of *P.* sp. "nyererei-like" through parallel speciation.

The *LWS* opsin gene is located on linkage group 13 in the anchored *P. nyererei* reference genome (fig. 5) corresponding to scaffold 177 of the original genome assembly by Brawand et al. (2014). Here, we found that the genomic region around this gene is highly differentiated between the sympatric species both at Makobe and at Python Island. The F_{ST} estimate of the 15-SNP window overlapping with the *LWS* gene was 0.84 (99.992th quantile) for the younger species pair, and 0.58 (99.434th quantile) for the older species pair. The region of

overlap between Makobe and Python HDRs spanned five genes including, besides *LWS* also the blue-sensitive opsin gene (*SWS-2 A*) and the violet-sensitive opsin gene (*SWS-2B*, fig. 5). Three 20-kb windows overlapped with this region. For two of the three 20-kb windows spanning the region, $f_d$nyer values were available and estimated to be very high (0.80 and 0.22, corresponding to the 99.7th and the 76.4th quantile) indicating high excess allele sharing between the two red-backed species. The third window contained too few informative SNPs to estimate $f_d$nyer but had a high $f_d$pund estimate (0.18, 80th quantile) indicating excess allele sharing between the two species with blue male nuptial coloration. The 20-kb window overlapping with the *LWS* gene showed strong clustering of the species by male coloration. The "color topology" had a weight of 0.84, whereas the otherwise predominant "geography topology" had a weight of only 0.06 in this window.

Six sites in the *LWS* gene showed parallel differences in allele frequency between the older species at Makobe Island and the younger species at Python Island. The blue (*SWS-2 A*) and violet vision (*SWS-2B*) opsin genes also each contained multiple sites with large parallel allele frequency shifts between the species at both islands, whereas most sites in intergenic regions were highly differentiated either only between the species at one island or at both islands but with nonparallel allele frequency shifts (fig. 5).

Discussion

In this study, we explore genomic patterns of differentiation in a cichlid species complex that we previously showed to consist of a younger and an older pair of sympatric sister species, each with one species with blue and the other with red-back male nuptial coloration. The younger pair evolved outside the distribution range of the older red species after hybridization between the latter and an allopatric population of the older blue species. We show that the genomic landscape of differentiation in both species pairs is heterogeneous with highly differentiated genomic regions spread across the entire genome even in the very young species pair. One-third of the genomic regions that are most strongly differentiated between the older species at Makobe Island are also strongly differentiated between the younger species at Python Island. However, two-thirds of the genomic regions that are differentiated between the younger species are not differentiated between the older species at all, consistent with an independent speciation event rather than "respeciation". In the following, we will discuss each of these results in more detail.

Widespread Genomic Regions of High Interspecific Differentiation

We find highly heterogeneous patterns of genome-wide differentiation between the species. Both at Makobe and at Python Island, almost 8% of all 15-SNP windows were assigned to the highly differentiated state, and these were clustered into >300 highly differentiated genomic regions in each species pair (table 1). In the younger species pair at Python Island, we detected slightly more but smaller differentiated genomic regions than in the much older original

species pair at Makobe Island. Some of these HDRs may be false positives, but we find a strong enrichment for additional selection statistics indicating that differentiation in most of these regions was driven by divergent selection. The large number of highly differentiated regions in the species at Python Island and significant genome-wide divergence even after exclusion of HDRs may seem surprising given that the species are estimated to be very young (probably <250 generations, Meier, Sousa, et al. 2017). Gene flow is likely ongoing as suggested by the quite high estimates of interspecific gene flow (Seehausen et al. 2008; Meier, Sousa, et al. 2017) and observations of intermediate phenotypes in nature (Seehausen 1997, 2009; Seehausen et al. 2008).

In young incipient species that diverge in the face of gene flow and are not fully isolated yet, restriction of gene flow is thought to be initially confined to very few genomic regions that experience strong divergent selection, a prediction that finds some support in other study systems (Kronforst et al. 2013; Nadeau et al. 2013; Marques et al. 2016, 2017). However, here we showed that when large genetic variation due to hybrid ancestry is present at functionally relevant genes, extensive genetic differentiation at many and genomically widespread loci may characterize even very young cases of sympatric speciation.

Divergent selection acting simultaneously on many regions across the genome may eventually lead to genome-wide reductions in gene flow (Barton 1983; Feder, Egan, et al. 2012; Feder, Gejji, et al. 2012; Flaxman et al. 2012; Feder et al. 2014). Genome-wide reproductive isolation in ecological speciation with gene flow may result from "genomic congealing" (Feder et al. 2014; Flaxman et al. 2014). In this model, during a period of divergent selection in the face of gene flow, low levels of linkage disequilibrium (LD) between many loci under divergent selection slowly build up, until positive feedback between LD and selection increases the effectiveness of selection, leading to a sudden cessation of gene flow (Feder et al. 2014; Flaxman et al. 2014).

Whereas models of genomic congealing require a very long phase of LD building preceding the sudden emergence of reproductive isolation, we find that despite the very young age of the sympatric species at Python Island, they already show genome-wide differentiation. This indicates that the slow phase involving the building of LD may not be needed if divergent selection acts on a hybrid population with segregating ancestry tracts containing multiple linked beneficial loci that evolved in response to selection in the source species. Hybrid populations will thus display large variation for the traits differing between the source species and linkage disequilibrium between some coadapted alleles is already present. This may allow a hybrid population to quickly respond to divergent selection in conditions resembling those under which the source species had diverged (here, e.g., to deeper water, red-shifted light spectra and female preferences for red male coloration under these conditions vs. more shallow water, less red-shifted light and female preference for blue male coloration). Hybridization may also cause allele combinations that lead to transgressive traits, novel trait combinations, or intrinsic and environment-mediated incompatibilities

(reviewed in Abbott et al. 2013). Admixture may thus importantly enhance the genetic variation allowing a population to speciate rapidly in response to divergent selection.

Some genomic regions of high differentiation are very large, the largest being over 2 Mb, and we find that on average HDRs fall into regions of lower recombination rate than the rest of the genome (fig. 3). When species diverge in the face of gene flow, regions of high differentiation are often associated with regions of low recombination (Butlin 2005; Turner et al. 2005; Hoffmann and Rieseberg 2008; Noor and Bennett 2009; Renaut et al. 2013; Roesti et al. 2013; Marques et al. 2016; Samuk et al. 2017). Theoretical models predict that regions of restricted recombination may facilitate species formation or persistence despite gene flow by protecting linkage disequilibrium among alleles conferring adaptation or barriers to gene flow (Rieseberg 2001; Bürger and Akerman 2011; Yeaman and Whitlock 2011). This may be particularly important in cases of hybrid parallel speciation, because coadapted alleles may remain linked in low recombination regions in the hybrid population and thus more quickly diverge again under parallel selection pressures. Divergent selection may also be easier to detect in regions of low recombination because the sweep signal is more extensive than in high recombination regions. Background selection is also expected to generate high F_{ST} values in regions of low recombination as it locally reduces the nucleotide diversity within species and thus inflates the relative divergence between the species (Charlesworth et al. 1993; Noor and Bennett 2009; Nachman and Payseur 2012; Cruickshank and Hahn 2014). However, this would be unexpected when incipient species are fully sympatric and exchange genes as is the case in the young species pairs we studied here (see also Cruickshank and Hahn 2014; Burri 2017). Indeed, we find overall low correlation between F_{ST} estimates for the different species pairs (supplementary fig. S11, Supplementary Material online) and strong additional signatures of selection in highly divergent regions (fig. 4), making a scenario where background selection is the main driver of divergence unlikely. Our study thus adds to the growing body of literature demonstrating an important role of variation in recombination rates in speciation with gene flow.

Evidence for Hybrid Parallel Speciation

More than 70% of the HDRs detected in the younger species pair at Python Island show no signatures of selection in the older species pair, suggesting that the younger species are not differentiated simply in a subset of the genomic differences that characterize the older species pair. It rather suggests that evolution of the younger species pair at Python Island involved divergence in many novel genomic regions, too. Given the very young age of this species pair, and the low proportions of alleles not shared with the older species pair, it is unlikely that de novo mutations underlie these novel HDRs. The alleles under divergent selection only in the younger pair may have been segregating neutrally in the older species pair or they may have been introduced into the younger species through introgression from a third species in the Mwanza Gulf. Given the very low proportions of private alleles, the former explanation is very likely.

Overall, the correlation is low between the differentiation landscapes of the two species pairs (supplementary fig. S11, Supplementary Material online), consistent with substantial independent evolution of these pairs. In general, the findings suggest that P. sp. "nyererei-like" and P. sp. "pundamilia-like" of Python Island and the north-western Mwanza Gulf do not merely represent reassembled genomes of the original P. nyererei and P. pundamilia (respeciation scenario sensu Gilman and Behm 2011). Instead, our data support the prediction of a hybrid parallel speciation hypothesis that the younger species of the Mwanza Gulf diverged largely independently from the older species pair outside of the Mwanza Gulf, whereby parallel selection targeted admixture-derived haplotypic variation in some genomic regions. The finding that shared HDRs have on an average higher mean F_{ST} values (supplementary fig. S4, Supplementary Material online) and a stronger enrichment of other selection statistics (fig. 4 and supplementary figs. S5, and S6, Supplementary Material online) than HDRs unique to either one of the species pairs, suggests that the haplotypes under strongest divergent selection in the original species pair were importantly involved in the divergence of the younger species pair.

Modest Levels of Genomic Parallelism despite Strong Phenotypic Parallelism

Even though the overlap of highly differentiated genomic regions detected in the species pairs of Python and Makobe Islands is higher than expected from random permutations (supplementary fig. S3, Supplementary Material online), less than a third of the HDRs in each species pair are shared between the two species pairs. This amount of genomic parallelism is unexpectedly modest given the extreme recency of the younger species pair (likely <250 generations, Meier, Sousa, et al. 2017). The probability of gene reuse under parallel selection pressures is expected to be very high when selection acts on shared standing variation (~80%, Conte et al. 2012; which may be slightly different at the level of genomic regions Roda et al. 2013). Due to the recent hybrid origin of both species sampled at Python Island and ongoing gene flow between all four species, as well as other additional cichlid species, the availability of genetic variation is unlikely to be a major constraint on parallel evolution in Pundamilia cichlids.

Modest parallelism at the genomic level despite strong phenotypic parallelism has also been found in many other young species/ecotype pairs, such as wave-swept and crab-predated Littorina ecotypes (Ravinet et al. 2016), or ecomorphs of Timema stick insects (Soria-Carrasco et al. 2014). Striking phenotypic parallelism despite modest levels of genomic parallelism may have multiple explanations and distinguishing among them is challenging. These explanations include: 1) The phenotypes may have a quantitative or complex genetic basis with multiple genes in the same pathway that can be tuned by selection with similar phenotypic output (Hoekstra and Nachman 2003; Pritchard et al. 2010; Boyle et al. 2017). 2) The few shared HDRs may underlie phenotypic parallelism, while the additional HDRs unique to each species pair may be under nonparallel divergent selection pressures, for example, due to different parasite regimes, diets, or water

clarity. 3) The observed phenotypic parallelism may be mainly due to phenotypic plasticity (Lande 2009; Pfennig et al. 2010). However, phenotypic plasticity can be ruled out as a cause of many species differences in *Pundamilia* (Seehausen 2009). There is evidence that the genetic basis underlying female mate choice and male nuptial coloration in the young *Pundamilia* species pair of Python Island is fairly simple and based on only a small number of physically unlinked Mendelian factors that may consist of single genes or multiple linked genes (Haesler and Seehausen 2005; Magalhaes and Seehausen 2010; Svensson et al. 2017). For most other traits that differ between the species at Makobe and at Python Island, the complexity of the genetic basis is unknown.

Some of the lack of genomic parallelism may be attributed to differences in the detection probabilities of signatures of selection in the older and younger species pair. The lower overall F_{ST} in the younger species pair may facilitate the detection of high F_{ST} outliers compared with the older species pair. Additionally, sorting of divergent ancestry blocks containing one or multiple sites under divergent selection in the younger species may generate a strong signal. If only a few adaptive haplotypes were present in the admixed ancestry of the younger species, the signal may resemble a hard sweep. The detection of such regions in the younger species pair may thus be easier than the detection of regions under long-term divergent selection (repeated soft sweeps) in the older species pair. On the other hand, sweeps may still be ongoing in the younger species pair and thus harder to detect. However, the different selection statistics we used in addition to F_{ST} have varied advantages and limitations in detecting different kinds of sweeps. As an example, because iHS is most sensitive to ongoing sweeps (Voight et al. 2006), it may have more power to detect recent sweeps in the younger species pair, whereas XP-EHH, which is most sensitive when the selected allele is near fixation (Sabeti et al. 2007), may have more power in the older species pair. The combination of these different statistics in our analysis should thus account for differences in selection signals in the two species pairs. Together, the array of selection statistics we used strongly support the presence of divergent selection in HDRs unique to one species pair and the absence of divergent selection in the same regions in the other species pair (fig. 4a and b).

Shared Highly Differentiated Genomic Regions Contain Candidates for Parallel Divergent Selection

Genomic regions that are highly differentiated between both species pairs are enriched for regions with high allele sharing between both red-backed species (high $f_d$nyer, fig. 4d) and (to a lesser extent) high allele sharing between both blue species (high $f_d$pund, fig. 4e). In many SHDRs, the species group more strongly by male nuptial coloration than by sampling location, against the genome-wide phylogenetic signatures (fig. 4f). The excess allele sharing between more distantly related species of similar phenotype in these genomic regions could be due to 1) higher *P. nyererei* ancestry proportions in *P.* sp. "nyererei-like" and higher *P. pundamilia* ancestry proportions in *P.* sp. "pundamilia-like", for instance due to differential sorting of the parental haplotypes, or 2) similar selection

pressures on alleles forming part of a shared pool of standing variation (i.e., already present in the Mwanza Gulf *Pundamilia* population before the admixture event), or 3) ongoing gene flow between the two red-backed species and between the two species with blue male nuptial coloration. More than half of the genomic regions that are highly differentiated between the species in both pairs show excess allele sharing between the two blue and the two red-backed species. This indicates that in these shared HDRs positive selection acts on *nyererei*-derived alleles in *P.* sp. "nyererei-like" and on *pundamilia*-derived alleles in *P.* sp. "pundamilia-like". It thus seems likely that *P. nyererei* alleles introduced into the ancestral *Pundamilia* of the north-western Mwanza Gulf in the inferred hybridization event (Meier, Sousa, et al. 2017) facilitated evolutionary response to divergent natural and sexual selection pressures between microhabitats. For example, red male nuptial coloration, zooplanktivory, and/or occupation of the deep water habitat are the aspects in which *P.* sp. "nyererei-like" resembles *P. nyererei*, and introduction of alleles related to these traits likely facilitated the evolution of *P.* sp. "nyererei-like".

We identified 41 candidate regions of parallel selection in the two species pairs, whereby in the younger species pair, divergent selection appears to act on the genetic variation derived from the hybridization event. They overlap with 78 genes with potential roles in sensory system, morphological, coloration, trophic, and immune system differences between the species. van Rijssel et al. (2018) showed that some of the parallel morphological differences between the species in the two species pairs are currently under disruptive selection in both species pairs. However, pinpointing the genes underlying these morphological traits requires further research. The candidate gene of parallel selection with most information available is the red-sensitive (*LWS*) opsin gene. The *LWS* opsin gene is under divergent selection between these species as the light environment is more red-shifted in deeper water where *P. nyererei* and *P.* sp. "nyererei-like" occur compared with the shallow-water habitat of *P. pundamilia* and *P.* sp. "pundamilia-like" (Seehausen et al. 2008). In addition, differences in color vision may modulate the female preferences for differently colored males (Kamijo et al. 2018) and thus contribute directly to reproductive isolation between the species. We demonstrate that the region around the *LWS* gene is highly differentiated between the species in both species pairs. The differences in allele frequencies between the species in the two pairs are remarkably parallel not just at the *LWS* gene but also at nearby genes, whereas they are not parallel in intergenic regions. In the entire region, genes are strongly enriched for SNPs with parallel allele frequency differences in the two species pairs, suggesting a role for parallel divergent selection on admixture-derived allelic variation in the large region containing three opsin genes (*LWS*, *SWS-2A*, *SWS-2B*) and associated regulatory regions.

Conclusions

Here, we studied a very young species pair that emerged from admixed ancestry of two older species that had evolved under parallel selection pressures. Despite their very recent origin

and ongoing gene flow, the genome-wide average differentiation between the fully sympatric younger species is significant and many genomic regions are highly differentiated. Some of those differentiated regions arose through divergent selection on haplotypes derived from the older species and are also among the most strongly differentiated regions in the older species pair. This finding implicates an important role of hybridization in the generation of genetic variation that facilitates rapid response to divergent natural and sexual selection through differential sorting of admixture-derived haplotypes with multiple coadapted alleles. In addition, many more genomic regions are divergent between the species in the younger species pair that do not show evidence of divergent selection in the older species pair. As supported by assortative mating between the younger and the older red-backed species in the laboratory, this finding shows that the younger species are not merely replicas of the original species, but instead are largely independent species. This independence may facilitate the eventual return to sympatry between the younger and the older species. Hybrid parallel speciation may thus be a process that facilitates the buildup of species richness in young adaptive radiations.

Materials and Methods

Samples and Sequencing

Wild males of *P. nyererei*, *P.* sp. "nyererei-like", *P. pundamilia*, and *P.* sp. "pundamilia-like" were collected with gill nets and angling at Makobe Island in the Speke Gulf and at Python Island in the north-western Mwanza Gulf in 2010 (fig. 1, Meier, Sousa, et al. 2017). The protocol for obtaining whole-genome sequencing data is detailed in Meier, Sousa, et al. (2017). Briefly, PCR-free libraries (Kozarewa et al. 2009) were produced from phenol-chloroform extracted DNA. To avoid any sequencing lane effects and to get an even read representation, all libraries were pooled after ligation of unique barcodes and sequenced together on four Illumina HiSeq 3000 lanes.

Local alignment against an anchored version of the *P. nyererei* reference genome (Brawand et al. 2014; Feulner et al. 2018) was performed with Bowtie 2 (Langmead and Salzberg 2012), and variant calling and genotyping was conducted with Haplotype Caller (GATK v. 3. 5, McKenna et al. 2010) following the best practice recommendations of GATK. To account for potential Illumina barcode switching issues (Sinha et al. 2017), we set the "contamination_fraction_to_filter" parameter in HaplotypeCaller to 0.1, to ignore 10% of the reads per allele of each genotype for genotype calling. In addition, we removed heterozygote genotypes with strongly unbalanced distribution of reads per allele by excluding genotypes failing a binomial test at a P value threshold of 0.001 (custom Python script "allelicBalance.py"). To avoid potential paralogous regions, we excluded sites with a mean depth per individual >1.5 times the interquartile range from the mean (i.e., 28.5, custom bash script "removeTooHighDepthSites.sh"), and sites with excess heterozygosity over all 16 individuals using the vcftools option "--hardy" and a P value cutoff of 0.01.

We also excluded sites based on the GATK variant annotations, including the fisher strand test, mapping quality, and the read position rank sum test (FS > 60, MQ < 40, MQRankSum < −12.5 and ReadPosRankSum < −8). Genotypes with less than ten reads were set as missing and sites with >50% missing data were excluded. If not otherwise indicated, all filtering steps were performed with vcftools v. 0.1.15 (Danecek et al. 2011).

Measures of Nucleotide Diversity and Selection

To study divergence and nucleotide diversity patterns along the genome, we used a sliding window approach on the whole-genome data set. All statistics were calculated for non-overlapping 20-kb windows. Absolute divergence (d_{XY}), nucleotide diversity (π), and the number of segregating sites (S) in each population were calculated with a modified version of a Python script by Martin et al. (2015). For both species pairs separately, we calculated the difference in nucleotide diversity between blue and red species, which is expected to be large if a selective sweep led to locally reduced diversity in one of the species. We also calculated Tajima's D using only sites without missing data in each species with the formula by Tajima (1989).

In order to identify regions of elevated relative divergence, we calculated pairwise locus-specific F_{ST} values between the species at each of the two islands. For each pairwise comparison, we created a data set of biallelic SNPs sequenced in all eight individuals of the compared population pair. We calculated within- and between-group variance components for each SNP with Arlequin v. 3.5.2 (Excoffier and Lischer 2010) using a locus-by-locus AMOVA including an individual level. We then calculated weighted average F_{ST} values for windows of 15 SNPs using the ratio of average variances across all SNPs in the sliding window as recommended by Bhatia et al. (2013), that is, as the mean variance between groups divided by the sum of the mean variance between groups, the mean variance within groups, and the mean individual variance. Windows of 15 SNPs provide a good balance between fine-scale genomic resolution and low stochastic variation by averaging over multiple variants (Malinsky et al. 2015). In addition, using windows of fixed number of variants as opposed to windows of fixed size provides the advantage that the variance is similar in all windows.

As additional measures of divergent selection, we used two extended haplotype length-based selection statistics (EHL), the cross-population extended haplotype homozygosity statistic (XP-EHH, Sabeti et al. 2007) and the integrated haplotype score (iHS, Voight et al. 2006). IHS compares the lengths of different haplotypes at the same genomic position within a population. It increases during ongoing sweeps as the sweeping haplotype is much longer than the other haplotypes. XP-EHH compares the lengths of haplotypes between populations and has most power to detect selection when the haplotype under selection is near fixation in one population and polymorphic in the other population. Background selection in regions of low recombination would lead to longer haplotypes in both species and would thus not increase XP-EHH (assuming conserved recombination rates, which is likely true

in this case). To compute these EHL statistics, we first phased the data set using fastPHASE (Scheet and Stephens 2006) assuming four clusters (K4). We predicted recombination distances between markers using the linkage map by Feulner et al. (2018), based on 1,597 SNPs and 212 F2 individuals from a *P. pundamilia* x *P.* sp. "red head" cross. First, we pruned the linkage map for outliers and markers that were <20 kb apart. Then we fitted a cubic smoothing spline to the physical and recombination distances using the R-function "smooth.spline" setting the smoothing parameter (spar) to 0.7 and predicted the recombination positions in cM for the genomic positions as the first derivative of the "predict.smooth.spline" function. We computed XP-EHH for both sympatric species pairs and iHS for each species with selscan v. 1.1.0 b (Szpiech and Hernandez 2014). Given that we do not know the ancestral and derived states of the SNPs, we used absolute iHS values. Similarly, as we do not distinguish in which species selection acted, we used absolute XP-EHH values. We normalized both statistics in allele frequency bins of 0.05 across all chromosomes combined as in Voight et al. (2006). Next, we computed the fraction of normalized absolute scores >2.0 in nonoverlapping 20-kb windows with norm v.1.1.0a (Szpiech and Hernandez 2014). A region with a high fraction of extreme scores is a better indicator of selective sweeps than single high scores (Voight et al. 2006). We identified "outlier windows" as top 5% windows with the highest fractions of extreme normalized scores among windows with similar number of SNPs (100 bins) using norm v.1.1.0a (Szpiech and Hernandez 2014).

Measures of Excess Allele Sharing

The gene flow statistic f_d is similar to D statistics (ABBA-BABA tests) but is more robust to low numbers of SNPs and thus more suited for gene flow estimates in small genomic regions (Martin et al. 2015). Like D, f_d is also used to measure excess allele sharing between two populations by comparing allele frequencies among four populations. The four populations are required to have a genome-wide average tree of the following topology: [(P1, P2), P3, O]. P3 is tested for excess allele sharing with P2 relative to its sister population P1, and O denotes an outgroup. In the absence of hybridization, P1 and P2 should both share equal numbers of alleles with P3. Excess allele sharing between P2 and P3 is usually interpreted as signature of gene flow but could also arise from parallel selection on shared ancestral polymorphisms. The outgroup is used to define the ancestral (A) and derived (B) state of the two alleles. The f_d statistic is then calculated as the difference in number of ABBA (P2 and P3 share the derived allele) and BABA (P1 and P3 share the derived allele) counts relative to the expected counts under a scenario of complete mixing between P2 and P3. For more details see Martin et al. (2015). High f_d estimates indicate excess allele sharing and thus likely gene flow between P2 and P3.

We calculated f_d using *Harpagochromis vittatus* from Lake Kivu as outgroup. Assuming that the species at Python Island are sister taxa, we tested for excess allele sharing between the two red-backed species, *P. nyererei* Makobe and *P.* sp. "nyererei-like" at Python ($f_d$nyer, P1 = *P.* sp. "pundamilia-like"

Python, P2 = *P.* sp. "nyererei-like", P3 = *P. nyererei*), and between the two blue species, *P. pundamilia* and *P.* sp. "pundamilia-like" ($f_d$pund, P1 = *P.* sp. "nyererei-like" Python, P2 = *P.* sp. "pundamilia-like" Python, P3 = *P. pundamilia* Makobe). To minimize false signatures of introgression due to low numbers of ancestry-informative sites, we calculated f_d statistics only for sliding windows where the sum of ABBA and BABA counts was at least five. As fixed ABBA or BABA patterns are rare, most sites fit the ABBA or BABA patterns only partially and are thus also only counted partially. As an example, if the allele frequencies are p_1=0.5, p_2=0, p_3=0.5, and p_o=0, it would count as 0.25 BABA pattern [$C_{BABA} = p_1 \times (1-p_2) \times p_3 \times (1-p_o)$]. Therefore, many sites partially fitting the ABBA or BABA pattern are needed to reach a total sum of five.

Identification of Highly Differentiated Genomic Regions

In order to identify genomic regions of high interspecific differentiation and to assess clustering of high F_{ST} windows, we applied a discrete state Hidden Markov Model (HMM) approach for the species pairs at Makobe and at Python Island separately.

We performed 10,000 coalescent simulations under the demographic model inferred to best explain the observed data (Meier, Sousa, et al. 2017) using fastsimcoal2 (Excoffier et al. 2013). We used the recombination rate of 2.3×10^{-8} per bp and generation (Feulner et al. 2018) and the mutation rate of 6.6×10^{-8} estimated by Recknagel et al. (2013). We simulated eight 20-kb sequences corresponding to four diploid individuals per species. We calculated pairwise F_{ST} estimates for both species pairs separately averaged across the first 15 SNPs like for the observed data. The distribution of the 10,000 weighted F_{ST} averages was then used as null distribution to compute z-scores for the observed 15 SNP F_{ST} averages. For each observed F_{ST} average, we computed the P value as the quantile of the null distribution. Next, we transformed P values into z-scores (i.e., normal scores) using the R-function "qnorm". The z-scores were approximately normally distributed indicating a good fit of the null model to the observed data.

An HMM analysis was then performed on the z-scores to assign each 15-SNP window to either the "standard" or "high differentiation". We fixed the means to 0 for the "standard state" and to 1.28 (P value of 0.1) for the "high differentiation state" and the SDs matching the empirical values. Following Hofer et al. (2012) and Marques et al. (2016), transition and emission probabilities were optimized with the Baum–Welch algorithm (6 runs with 1,000 iterations each, Baum et al. 1970) and the most likely sequence of states given the observed z-scores was inferred with the Viterbi algorithm (Viterbi 1967). As in Marques et al. (2016), we concatenated the chromosomes to increase the information for parameter estimation. We did not find spurious regions of high differentiation resulting from concatenation of the chromosomes, that is, no erroneous islands extending over multiple chromosomes.

To assess clustering of highly differentiated windows, we concatenated directly adjacent windows assigned to the "high differentiation" state to "highly differentiated genomic

regions" (HDR). Genomic regions where HDRs from Makobe and from Python Island overlapped were termed shared HDRs. We used a block-permutation approach to assess if the number and length of these shared HDRs differed from random expectations. We permuted the HMM states for each species pair separately in blocks of 100 windows and recalculated the number and mean lengths of shared HDRs. A block size of 100 windows exceeds the length of 98% of HDRs allowing us to account for the clustered distribution of highly differentiated windows. The permutation procedure was repeated 100 times to get a random distribution of overlap statistics.

Genome-wide average differentiation between the sympatric species outside of the highly divergent genomic regions was assessed by computing mean F_{ST} values from SNPs without missing data that were pruned for high linkage disequilibrium ($r^2 > 0.5$) with plink v.1.07 (using –indep-pairwise 50 10 0.5, Purcell et al. 2007) and vcftools v. 0.1.15 (Danecek et al. 2011) using a custom script (ldPruning.sh). Average F_{ST} values between the species were computed with an ANOVA performed in Arlequin v. 3.5.2 (Excoffier and Lischer 2010). We computed confidence intervals from 20,000 bootstraps to test if the differentiation was significant.

In order to test the robustness of the detection of HDRs, we additionally used another approach to identify HDRs. We computed window averages for nonoverlapping 20-kb windows and for 50-kb windows. We identified the windows with F_{ST} estimates above the 90th quantile and concatenated consecutive windows. To avoid splitting of HDRs due to short stochastic dips in F_{ST}, we ignored single windows with F_{ST} estimates below the threshold within runs of highly differentiated windows. The HDRs were identified both for the Makobe and the Python species pairs and compared with the HDRs identified with the HMM approach based on 15-SNP windows.

To test for genome-wide correlation of differentiation between the two species pairs, we compared F_{ST} estimates of the older species pair at Makobe Island with F_{ST} estimates of the younger species pair at Python Island in the same genomic regions. In addition, we used restriction-site associated DNA (RAD) sequencing data from Meier, Sousa, et al. (2017) to compare the extent of correlation between the species pairs to the correlation of differentiation with another *Pundamilia* population from Luanso Island. Luanso Island is ~7.5 km away from Python Island and ~37.5 km from Makobe Island. It harbors a phenotypically polymorphic but undifferentiated population of *Pundamilia* cichlids and is thus ideal to assess the expected level of correlation without divergent selection (e.g., due to background selection). In addition, we used RAD data from a second sampling location of the younger species pair, Kissenda Island, which is 9.6 km north of Python Island. We used 20 individuals of *P. nyererei* and 19 individuals of *P. pundamilia*, all sampled at Makobe Island. We included 16 or 22 *P.* sp. "nyererei-like" and 26 or 22 individuals of *P.* sp. "pundamilia-like" from Kissenda Island or from Python Island, respectively. Of the *Pundamilia* population at Luanso Island, we used 53 individuals. We used only sites sequenced in at least ten individuals per population or species with at least ten reads per individual. Otherwise, we applied the same filtering steps as detailed in Meier, Sousa, et al. (2017). F_{ST} values were estimated for all species/population pairs using the site-by-site ANOVA implemented in Arlequin v. 3.5.2 (Excoffier and Lischer 2010). We computed the ratio of average variances (Bhatia et al. 2013) across all SNPs in 20-kb nonoverlapping windows both for the RAD data set and the whole-genome data set. Windows with less than three SNPs were discarded. Correlation between F_{ST} estimates of different pairs was assessed with Pearson's correlation using the R function "cor.test". To further assess genomic parallelism between the pairs, we extracted all 20-kb windows with a RAD F_{ST} estimate >0.2 for the Makobe species pair. We then tested if these windows also exhibit higher RAD F_{ST} estimates than null expectations from 1,000 permuted data sets in the younger species pair sampled at Python and Kissenda Island and in the control comparison of each Python species with the Luanso population.

Parallelism at the Genomic Level

We tested for parallelism at the genomic level by assessing the predominant tree topologies in 20-kb windows using TWISST (Martin and Van Belleghem 2017). First, we computed for each 20-kb window a maximum likelihood tree of all 16 individuals using raxmlHPC-AVX (v. 8.2.4, Stamatakis 2014) and raxmlWindows.py (https://github.com/simonhmartin/genomics_general, last accessed March 14, 2018). Next, we used TWISST to iteratively subsample all combinations of a single individual per species (option "complete") and compute the proportion of subtrees matching each of the possible species topologies. With four species, three unrooted tree topologies are possible: the "geography topology", where species of Python Island cluster together and species of Makobe Island cluster together, the "color topology", where *P. nyererei* and *P.* sp. "nyererei-like" cluster together and *P. pundamilia* and *P.* sp. "pundamilia-like" cluster together, and the "random topology", where *P. pundamilia* clusters with *P.* sp. "nyererei-like" and *P. nyererei* clusters with *P.* sp. "pundamilia-like". High weights of the "color topology" in a genomic region may be due to parallel selection in the two species pairs.

We determined candidate regions of parallel selection as 20-kb windows in shared highly divergent regions (SHDRs) with a weight of at least 66% for the "color topology". A weight of 66% means that the topology is twice as well supported than the two other topologies together. For these windows, we identified overlapping genes from the NCBI *P. nyererei* RefSeq annotation release 101 (https://www.ncbi.nlm.nih.gov/genome/annotation_euk/Pundamilia_nyererei/101/#BuildInfo, last accessed March 15, 2018) lifted to the new anchored version of the *P. nyererei* reference genome (Brawand et al. 2014; Feulner et al. 2018) using the UCSC tool liftOver (Hinrichs et al. 2006) and custom chain files. We searched OrthoDB (Zdobnov et al. 2017) for orthologs of these candidate genes in *Danio rerio*, *Mus musculus*, *Rattus norvegicus*, and *Homo sapiens*, extracted associated functional annotation for these orthologs and assessed their functional

relevance in light of the natural history of *Pundamilia* speciation.

Supplementary Material

Supplementary data are available at *Molecular Biology and Evolution* online. Lists and plots of candidate regions of parallel selection on admixture-derived variation are available on the Dryad Digital Repository: https://doi.org/10.5061/dryad.8p6m07g.

Acknowledgments

Author Contributions

J.I.M. and O.S. designed the study; J.I.M. performed the analyses with assistance of the other authors; D.A.M. explored the gene functions; J.I.M. wrote the manuscript with contributions from O.S., L.E., C.E.W., and D.A.M.

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(2):229–246.

Barton N. 1983. Multilocus clines. *Evolution* 37(3):454–471.

Baum LE, Petrie T. 1966. Statistical inference for probabilistic functions of finite state markov chains. *Ann Math Stat.* 37(6):1554–1563.

Baum LE, Petrie T, Soules G, Weiss N. 1970. A maximization technique occurring in the statistical analysis of probabilistic functions of markov chains. *Ann Mat Stat.* 41(1):164–171.

Bhatia G, Patterson N, Sankararaman S, Price AL. 2013. Estimating and interpreting F_{ST}: the impact of rare variants. *Genome Res.* 23(9):1514–1521.

Bouton N, Seehausen O, Alphen JJM. 1997. Resource partitioning among rock-dwelling haplochromines (Pisces: Cichlidae) from Lake Victoria. *Ecol Freshwater Fish* 6(4):225–240.

Boyle EA, Li YI, Pritchard JK. 2017. An expanded view of complex traits: from polygenic to omnigenic. *Cell* 169(7):1177–1186.

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.

Bürger R, Akerman A. 2011. The effects of linkage and gene flow on local adaptation: a two-locus continent–island model. *Theor Pop Gen.* 80(4):272–288.

Burri R. 2017. Interpreting differentiation landscapes in the light of long-term linked selection. *Evol Lett.* 1(3):118–131.

Burri R, Nater A, Kawakami T, Mugal CF, Olason PI, Smeds L, Suh A, Dutoit L, Bureš S, Garamszegi LZ, et al. 2015. Linked selection and recombination rate variation drive the evolution of the genomic

landscape of differentiation across the speciation continuum of ficedula flycatchers. *Genome Res.* 25(11):1656–1665.

Butlin RK. 2005. Recombination and speciation. *Mol Ecol.* 14(9):2621–2635.

Carleton K, Parry J, Bowmaker J, Hunt D, Seehausen O. 2005. Color vision and speciation in Lake Victoria cichlids of the genus *Pundamilia*. *Mol Ecol.* 14(14):4341.

Charlesworth B, Morgan MT, Charlesworth D. 1993. The effect of deleterious mutations on neutral molecular variation. *Genetics* 134(4):1289–1303.

Conte GL, Arnegard ME, Peichel CL, Schluter D. 2012. The probability of genetic parallelism and convergence in natural populations. *Proc Biol Sci B* 279(1749):5039–5047.

Coyne JA, Orr HA. 2004. Speciation. Sunderland (MA): Sinauer Associates.

Cruickshank TE, Hahn MW. 2014. Reanalysis suggests that genomic islands of speciation are due to reduced diversity, not reduced gene flow. *Mol Ecol.* 23(13):3133–3157.

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(15):2156–2158.

Excoffier L, Dupanloup I, Huerta-Sánchez E, Sousa VC, Foll M. 2013. Robust demographic inference from genomic and SNP data. *PLoS Genet.* 9(10):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 Resour.* 10(3):564–567.

Feder JL, Egan SP, Nosil P. 2012. The genomics of speciation-with-gene-flow. *Trends Genet.* 28:342–350.

Feder JL, Gejji R, Yeaman S, Nosil P. 2012. Establishment of new mutations under divergence and genome hitchhiking. *Philos Trans R Soc B* 367(1587):461–474.

Feder JL, Nosil P. 2010. The efficacy of divergence hitchhiking in generating genomic islands during ecological speciation. *Evolution* 64(6):1729–1747.

Feder JL, Nosil P, Wacholder AC, Egan SP, Berlocher SH, Flaxman SM. 2014. Genome-wide congealing and rapid transitions across the speciation continuum during speciation with gene flow. *J Hered.* 105(S1):810–820.

Feulner PGD, Schwarzer J, Haesler MP, Meier JI, Seehausen O. 2018. A dense linkage map of Lake Victoria cichlids improved the *Pundamilia* genome assembly and revealed a major qtl for sex-determination. bioRxiv. doi: https://doi.org/10.1101/275388.

Flaxman S, Feder J, Nosil P. 2012. Spatially explicit models of divergence and genome hitchhiking. *J Evol Biol.* 25(12):2633–2650.

Flaxman SM, Wacholder AC, Feder JL, Nosil P. 2014. Theoretical models of the influence of genomic architecture on the dynamics of speciation. *Mol Ecol.* 23(16):4074–4088.

Gilman RT, Behm JE. 2011. Hybridization, species collapse, and species reemergence after disturbance to premating mechanisms of reproductive isolation. *Evolution* 65(9):2592–2605.

Grant PR, Grant BR. 1992. Hybridization of bird species. *Science* 256(5054):193–197.

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.

Hinrichs AS, Karolchik D, Baertsch R, Barber GP, Bejerano G, Clawson H, Diekhans M, Furey TS, Harte RA, Hsu F, et al. 2006. The UCSC genome browser database: update 2006. *Nucl Acids Res.* 34:D590–D598.

Hoekstra HE, Nachman MW. 2003. Different genes underlie adaptive melanism in different populations of rock pocket mice. *Mol Ecol.* 12(5):1185–1194.

Hofer T, Foll M, Excoffier L. 2012. Evolutionary forces shaping genomic islands of population differentiation in humans. *BMC Genomics* 13(1):107.

Hoffmann AA, Rieseberg LH. 2008. Revisiting the impact of inversions in evolution: from population genetic markers to drivers

of adaptive shifts and speciation? *Annu Rev Ecol Evol Syst.* 39:21–42.

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 B* 365(1547):1735–1747.

Johnson TC, Kelts K, Odada E. 2000. The holocene history of Lake Victoria. *Ambio* 29(1):2–11.

Kamijo M, Kawamura M, Fukamachi S. 2018. Loss of red opsin genes relaxes sexual isolation between skin-colour variants of medaka. *Behav Process.* 150:25–28.

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(11):2848–2863.

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(4):291–295.

Kronforst MR, Hansen ME, Crawford NG, Gallant JR, Zhang W, Kulathinal RJ, Kapan DD, Mullen SP. 2013. Hybridization reveals the evolving genomic architecture of speciation. *Cell Rep.* 5(3):666–677.

Lande R. 2009. Adaptation to an extraordinary environment by evolution of phenotypic plasticity and genetic assimilation. *J Evol Biol.* 22(7):1435–1446.

Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nat Methods* 9(4):357–U354.

Maan ME, Hofker KD, van Alphen JJM, Seehausen O. 2006. Sensory drive in cichlid speciation. *Am Nat.* 167(6):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(1556):2445–2452.

Maan ME, Van Rooijen AMC, Van Alphen JJM, Seehausen O. 2008. Parasite-mediated sexual selection and species divergence in Lake Victoria cichlid fish. *Biol J Linn Soc.* 94(1):53–60.

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(2):260–274.

Magalhaes IS, Seehausen O. 2010. Genetics of male nuptial colour divergence between sympatric sister species of a Lake Victoria cichlid fish. *J Evol Biol.* 23(5):914–924.

Malinsky M, Challis RJ, Tyers AM, Schiffels S, Terai Y, Ngatunga BP, Miska EA, Durbin R, Genner MJ, Turner GF. 2015. Genomic islands of speciation separate cichlid ecomorphs in an East African crater lake. *Science* 350(6267):1493–1498.

Marques DA, Lucek K, Haesler MP, Feller AF, Meier JI, Wagner CE, Excoffier L, Seehausen O. 2017. Genomic landscape of early ecological speciation initiated by selection on nuptial colour. *Mol Ecol.* 26(1):7–24.

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(2):e1005887.

Martin SH, Dasmahapatra KK, Nadeau NJ, Salazar C, Walters JR, Simpson F, Blaxter M, Manica A, Mallet J, Jiggins CD. 2013. Genome-wide evidence for speciation with gene flow in *Heliconius* butterflies. *Genome Res.* 23(11):1817–1828.

Martin SH, Davey JW, Jiggins CD. 2015. Evaluating the use of ABBA-BABA statistics to locate introgressed loci. *Mol Biol Evol.* 32(1):244–257.

Martin SH, Jiggins CD. 2017. Interpreting the genomic landscape of introgression. *Curr Opin Genet Dev.* 47:69–74.

Martin SH, Van Belleghem SM. 2017. Exploring evolutionary relationships across the genome using topology weighting. *Genetics* 206(1):429–438.

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA. 2010. The genome analysis toolkit: a mapreduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* 20(9):1297–1303.

Meier JI, Marques DA, Mwaiko S, Wagner CE, Excoffier L, Seehausen O. 2017. Ancient hybridization fuels rapid cichlid fish adaptive radiations. *Nat Commun.* 8:14363.

Meier JI, Sousa VC, Marques DA, Selz OM, Wagner CE, Excoffier L, Seehausen O. 2017. Demographic modelling with whole-genome data reveals parallel origin of similar *Pundamilia* cichlid species after hybridization. *Mol Ecol.* 26:123–141.

Nachman MW, Payseur BA. 2012. Recombination rate variation and speciation: theoretical predictions and empirical results from rabbits and mice. *Philos Trans R Soc B* 367(1587):409–421.

Nadeau NJ, Martin SH, Kozak KM, Salazar C, Dasmahapatra KK, Davey JW, Baxter SW, Blaxter ML, Mallet J, Jiggins CD. 2013. Genome-wide patterns of divergence and gene flow across a butterfly radiation. *Mol Ecol.* 22(3):814–826.

Noor MAF, Bennett SM. 2009. Islands of speciation or mirages in the desert? Examining the role of restricted recombination in maintaining species. *Heredity* 103(6):439–444.

Nosil P, Harmon LJ, Seehausen O. 2009. Ecological explanations for (incomplete) speciation. *Trends Ecol Evol.* 24(3):145–156.

Pfennig DW, Wund MA, Snell-Rood EC, Cruickshank T, Schlichting CD, Moczek AP. 2010. Phenotypic plasticity's impacts on diversification and speciation. *Trends Ecol Evol.* 25(8):459–467.

Pritchard JK, Pickrell JK, Coop G. 2010. The genetics of human adaptation: hard sweeps, soft sweeps, and polygenic adaptation. *Curr Biol.* 20(4):R208–R215.

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ, Sham PC. 2007. Plink: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet.* 81(3):559–575.

Ravinet M, Westram A, Johannesson K, Butlin R, Andre C, Panova M. 2016. Shared and nonshared genomic divergence in parallel ecotypes of *Littorina saxatilis* at a local scale. *Mol Ecol.* 25(1):287–305.

Recknagel H, Elmer KR, Meyer A. 2013. A hybrid genetic linkage map of two ecologically and morphologically divergent midas cichlid fishes *Amphilophus* spp.) obtained by massively parallel DNA sequencing (ddradseq). *G3 (Bethesda)* 3:65–74.

Renaut S, Grassa C, Yeaman S, Moyers B, Lai Z, Kane N, Bowers J, Burke J, Rieseberg L. 2013. Genomic islands of divergence are not affected by geography of speciation in sunflowers. *Nat Commun.* 4:1827.

Riesch R, Muschick M, Lindtke D, Villoutreix R, Comeault AA, Farkas TE, Lucek K, Hellen E, Soria-Carrasco V, Dennis SR, et al. 2017. Transitions between phases of genomic differentiation during stick-insect speciation. *Nat Ecol Evol.* 1(4):0082.

Rieseberg LH. 2001. Chromosomal rearrangements and speciation. *Trends Ecol Evol.* 16(7):351–358.

Roda F, Liu H, Wilkinson MJ, Walter GM, James ME, Bernal DM, Melo MC, Lowe A, Rieseberg LH, Prentis P, Ortiz-Barrientos D. 2013. Convergence and divergence during the adaptation to similar environments by an australian groundsel. *Evolution* 67(9):2515–2529.

Roesti M, Moser D, Berner D. 2013. Recombination in the threespine stickleback genome—patterns and consequences. *Mol Ecol.* 22(11):3014–3027.

Rundle HD, Nagel L, Boughman JW, Schluter D. 2000. Natural selection and parallel speciation in sympatric sticklebacks. *Science* 287(5451):306–308.

Sabeti PC, Varilly P, Fry B, Lohmueller J, Hostetter E, Cotsapas C, Xie XH, Byrne EH, McCarroll SA, Gaudet R, Consortium IH, et al. 2007. Genome-wide detection and characterization of positive selection in human populations. *Nature* 449:913–U912.

Samuk K, Owens GL, Delmore KE, Miller SE, Rennison DJ, Schluter D. 2017. Gene flow and selection interact to promote adaptive divergence in regions of low recombination. *Mol Ecol.* 26(17):4378–4390.

Scheet P, Stephens M. 2006. A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *Am J Hum Genet.* 78(4):629–644.

Schluter D, Nagel LM. 1995. Parallel speciation by natural selection. *Am Nat.* 146(2):292–301.

Seehausen O. 1996. Lake Victoria rock cichlids. Taxonomy, ecology and distribution. Zevenhuizen (the Netherlands): Verduijn Cichlids.

Seehausen O. 1997. Distribution of and reproductive isolation among color morphs of a rock-dwelling Lake Victoria cichlid (*Haplochromis nyererei*) (vol 5, pg 195, 1996). *Ecol Freshwater Fish* 6(1):59–66.

Seehausen O. 2002. Patterns in fish radiation are compatible with Pleistocene desiccation of Lake Victoria and 14 600 year history for its cichlid species flock. *Proc Biol Sci B* 269(1490):491–497.

Seehausen O. 2004. Hybridization and adaptive radiation. *Trends Ecol Evol.* 19(4):198–207.

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, 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, 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(7213):620–626.

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):1–8.

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(3):612–626.

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 Biol Sci B* 283(1830):20160172.

Sinha R, Stanley G, Gulati GS, Ezran C, Travaglini KJ, Wei E, Chan CKF, Nabhan AN, Su T, Morganti RM, et al. 2017. Index switching causes "spreading-of-signal" among multiplexed samples in illumina hiseq 4000 DNA sequencing. bioRxiv, doi: 10.1101/125724.

Soria-Carrasco V, Gompert Z, Comeault AA, Farkas TE, Parchman TL, Johnston JS, Buerkle CA, Feder JL, Bast J, Schwander T, et al. 2014. Stick insect genomes reveal natural selection's role in parallel speciation. *Science* 344(6185):738–742.

Stager JC, Johnson TC. 2008. The late Pleistocene desiccation of Lake Victoria and the origin of its endemic biota. *Hydrobiologia* 596(1):5–16.

Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30(9):1312–1313.

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(1505):2861–2870.

Svensson O, Woodhouse K, van Oosterhout C, Smith A, Turner GF, Seehausen O. 2017. The genetics of mate preferences in hybrids between two young and sympatric Lake Victoria cichlid species. *Proc R Soc B* 284(1849):20162332.

Szpiech ZA, Hernandez RD. 2014. Selscan: an efficient multithreaded program to perform EHH-based scans for positive selection. *Mol Biol Evol.* 31(10):2824–2827.

Tajima F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* 123(3):585–595.

Turner TL, Hahn MW, Nuzhdin SV. 2005. Genomic islands of speciation in *Anopheles gambiae*. *PLoS Biol.* 3(9):e285–1578.

van Rijssel JC, Moser FN, Frei D, Seehausen O. 2018. Prevalence of disruptive selection predicts extent of species differentiation in Lake Victoria cichlids. *Proc Biol Sci.* 285(1871):20172630.

Via S. 2012. Divergence hitchhiking and the spread of genomic isolation during ecological speciation-with-gene-flow. *Philos Trans R Soc B* 367(1587):451–460.

Via S, West J. 2008. The genetic mosaic suggests a new role for hitchhiking in ecological speciation. *Mol Ecol.* 17(19):4334–4345.

Viterbi A. 1967. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Trans Informat Theory* 13(2):260–269.

Voight BF, Kudaravalli S, Wen XQ, Pritchard JK. 2006. A map of recent positive selection in the human genome. *PLoS Biol.* 4(3):e72–458.

Wu CI. 2001. The genic view of the process of speciation. *J Evol Biol.* 14(6):851–865.

Yeaman S, Whitlock MC. 2011. The genetic architecture of adaptation under migration-selection balance. *Evolution* 65(7):1897–1911.

Zdobnov EM, Tegenfeldt F, Kuznetsov D, Waterhouse RM, Simão FA, Ioannidis P, Seppey M, Loetscher A, Kriventseva EV. 2017. OrthoDB v9.1: cataloging evolutionary and functional annotations for animal, fungal, plant, archaeal, bacterial and viral orthologs. *Nucleic Acids Res.* 45(D1):D744–D749.