Ancestral Hybridization Facilitated Species Diversification in the Lake Malawi Cichlid Fish Adaptive Radiation

Hannes Svardal (ORCID) ,*[1,2,3,4] Fu Xiang Quah,[3] Milan Malinsky (ORCID) ,[5] Benjamin P. Ngatunga,[6] Eric A. Miska,[2,3,7] Walter Salzburger,[5] Martin J. Genner,[8] George F. Turner,[9] and Richard Durbin (ORCID) *,[2,3]

[1]Department of Biology, University of Antwerp, Antwerp, Belgium

[2]Department of Genetics, University of Cambridge, Cambridge, United Kingdom

[3]Wellcome Sanger Institute, Hinxton, United Kingdom

[4]Naturalis Biodiversity Center, Leiden, The Netherlands

[5]Zoological Institute, University of Basel, Basel, Switzerland

[6]Tanzania Fisheries Research Institute, Dar es Salaam, Tanzania

[7]Gurdon Institute, University of Cambridge, Cambridge, United Kingdom

[8]School of Biological Sciences, University of Bristol, Bristol, United Kingdom

[9]School of Biological Sciences, Bangor University, Bangor, United Kingdom

*Corresponding authors: E-mails: hannes.svardal@uantwerpen.be; rd109@cam.ac.uk.
Associate editor: Rasmus Nielsen

Abstract

The adaptive radiation of cichlid fishes in East African Lake Malawi encompasses over 500 species that are believed to have evolved within the last 800,000 years from a common founder population. It has been proposed that hybridization between ancestral lineages can provide the genetic raw material to fuel such exceptionally high diversification rates, and evidence for this has recently been presented for the Lake Victoria region cichlid superflock. Here, we report that Lake Malawi cichlid genomes also show evidence of hybridization between two lineages that split 3–4 Ma, today represented by Lake Victoria cichlids and the riverine *Astatotilapia* sp. "ruaha blue." The two ancestries in Malawi cichlid genomes are present in large blocks of several kilobases, but there is little variation in this pattern between Malawi cichlid species, suggesting that the large-scale mosaic structure of the genomes was largely established prior to the radiation. Nevertheless, tens of thousands of polymorphic variants apparently derived from the hybridization are interspersed in the genomes. These loci show a striking excess of differentiation across ecological subgroups in the Lake Malawi cichlid assemblage, and parental alleles sort differentially into benthic and pelagic Malawi cichlid lineages, consistent with strong differential selection on these loci during species divergence. Furthermore, these loci are enriched for genes involved in immune response and vision, including opsin genes previously identified as important for speciation. Our results reinforce the role of ancestral hybridization in explosive diversification by demonstrating its significance in one of the largest recent vertebrate adaptive radiations.

Key words: adaptive radiation, gene flow, hybrid swarm, cichlid fish.

Introduction

Adaptive radiation, the rapid diversification of one or a few ancestral lineages into a number of species that occupy a range of ecological niches, is thought to be responsible for a considerable fraction of extant biodiversity (Berner and Salzburger 2015). A growing body of research investigates the intrinsic and extrinsic factors that may facilitate adaptive radiation (reviewed in Seehausen 2015). This may be addressed by correlating ecological factors and organismal features with the propensity for adaptive diversification across taxa (Wagner et al. 2012), but the ultimate goal must be to gain a functional understanding of the genetic and ecological mechanisms at play.

The role of hybridization in adaptive radiation and speciation has been debated (Mallet 2007; Abbott et al. 2013; Schumer et al. 2014). Although the cessation of genetic exchange between sister lineages is often seen as a prerequisite for speciation, it has been proposed that both hybridization into the ancestral lineage as well as genetic exchange between diverging lineages within a radiation could favor adaptive diversification (Seehausen 2004, 2013; Marques et al. 2019). With genome-wide sequencing data becoming increasingly

Article

accessible, the latter scenario of genetic exchange between diverging lineages has been shown to be widespread in plants and animals (Mallet et al. 2016; Novikova et al. 2016; Svardal et al. 2017), especially in young adaptive radiations (Dowling and DeMarais 1993; Heliconius Genome Consortium 2012; Lamichhaney et al. 2015; Stryjewski and Sorenson 2017; Edelman et al. 2019; Kozak et al. 2018; Malinsky et al. 2018), and links to adaptation have been made (Heliconius Genome Consortium 2012; Richards and Martin 2017; Malinsky et al. 2018). However, the generality and evolutionary importance of hybridization in the seeding population of an adaptive radiation, that is, the ancestral lineage before initial divergence, is less well understood.

Hybridization in the common ancestor of an adaptive radiation could promote adaptive radiation in several ways (reviewed in Seehausen 2004; Marques et al. 2019). Hybrids often exhibit novel or extreme characters, known as transgressive segregation (Rieseberg et al. 1999). In particular, epistatic interactions of alleles or—at a higher level—genetic modules from two or more parental backgrounds can lead to novel phenotypes (Dittrich-Reed and Fitzpatrick 2013) that could confer a fitness advantage in previously underutilized ecological niches or provide a target for sexual selection. However, even in the absence of interactions between loci, hybridization is expected to increase the genetic variance of a quantitative trait, if alleles of different effect sizes and directions segregate in the two parental populations. This is true even if the parental lineages were (independently) under stabilizing selection for the same trait optimum. The heritable genetic variability produced in this way could facilitate adaptive divergence. Furthermore, genetic incompatibility, that is, negative epistasis between loci fixed for different alleles in the parental lineages, could potentially provide an axis of divergent selection (Schumer et al. 2015) that couples with and reinforces natural divergent selection, but whether this actually happens in nature and—if it does—how common it is, is largely an open question.

For a number of taxa, empirical evidence has been presented that ancestral hybridization can occur prior to adaptive radiation (Barrier et al. 1999; Meier et al. 2017) and can generate phenotypic novelty (Richards and Martin 2017; Stryjewski and Sorenson 2017). A group of species that has received considerable attention with respect to their repeated rapid adaptive diversification and the role of hybridization in this process are haplochromine cichlids, a tribe of percomorph fish widely distributed in eastern Africa including the species of the adaptive radiations in the Lake Victoria region, in Lake Malawi, and a subset of the species of the adaptive radiation of cichlid fishes in Lake Tanganyika (Kocher 2004; Salzburger 2018). Meier et al. (2017) demonstrated that the haplochromine cichlid fishes collectively known as the Lake Victoria region superflock (LVRS) emerged from a hybrid ancestor, that loci highly divergent across LVRS lineages are enriched for alleles differentiated between the two ancestral lineages, and that ecologically divergent opsin alleles originate from the two ancestors. Furthermore, hybridization was also reported between ancestral lineages of Lake Tanganyika cichlids (Irisarri et al. 2018) as well as between

some of the main lineages (tribes) in this lake (Meyer et al. 2016). Ancient hybridization has also been suggested to play a role in the Lake Malawi cichlid adaptive radiation (Joyce et al. 2011; Genner and Turner 2012), but the suggested events were based on mitochondrial evidence, which was not supported by recent whole-genome sequencing analyses (Malinsky et al. 2018).

Here we reinvestigate the occurrence and role of ancestral hybridization in the Lake Malawi cichlid adaptive radiation by analyzing recently published and newly obtained whole-genome sequences of haplochromine cichlids from Lake Malawi, the LVRS, as well as other East African lakes and rivers (fig. 1 and supplementary table 1, Supplementary Material online) (Brawand et al. 2014; McGee et al. 2016; Malinsky et al. 2018). We find strong evidence for ancient hybridization at the base of the Malawi cichlid radiation, between a lineage related to the LVRS and a riverine lineage previously shown to carry a Malawi-like mitochondrial haplotype (Genner et al. 2015). Furthermore, we show that the majority of the genetic variation produced by this hybridization event has been lost in Malawi by the relatively rapid fixation at most loci of one or the other parental contribution in the ancestral population of Malawi cichlids. The genetic variation created by this hybridization event that still persists in Malawi shows elevated divergence across the early splits in the radiation, beyond its effect of increasing minor allele frequencies, and is enriched in gene ontologies related to vision and the immune system.

Results

Sampling, Sequencing, and Variant Detection

We obtained samples for representatives of the haplochromine cichlid adaptive radiations in the Lake Victoria region, from Lake Malawi, and from riverine and lacustrine haplochromines that have been proposed to be close relatives of these radiations (fig. 1a and supplementary table 1, Supplementary Material online, Materials and Methods) (Genner et al. 2015; Malinsky et al. 2018). The Lake Malawi sampling comprised two samples for each of the previously defined major ecomorphological groups (only one sample for utaka, supplementary table 1, Supplementary Material online) (Malinsky et al. 2018). Illumina short-read sequence data were partly obtained from previous studies (Brawand et al. 2014; McGee et al. 2016; Malinsky et al. 2018) and partly sequenced for this study (supplementary table 1, Supplementary Material online). DNA libraries were created and whole-genome sequenced on an Illumina HiSeq platform to individual coverages between 7.9- and 19.2-fold (median 14.3-fold). Reads were aligned to the Nile tilapia reference genome Orenil. 1.1 (GCA_000188235.2) (Brawand et al. 2014) and variants called. After filtering, 19 million biallelic single nucleotide polymorphism (SNP) variants were kept for further analysis (Materials and Methods).

Gene Flow into the Ancestral Population of the Lake Malawi Cichlid Flock

To provide an overview of genetic relationships between all samples, we first built a neighbor-joining (NJ) tree from
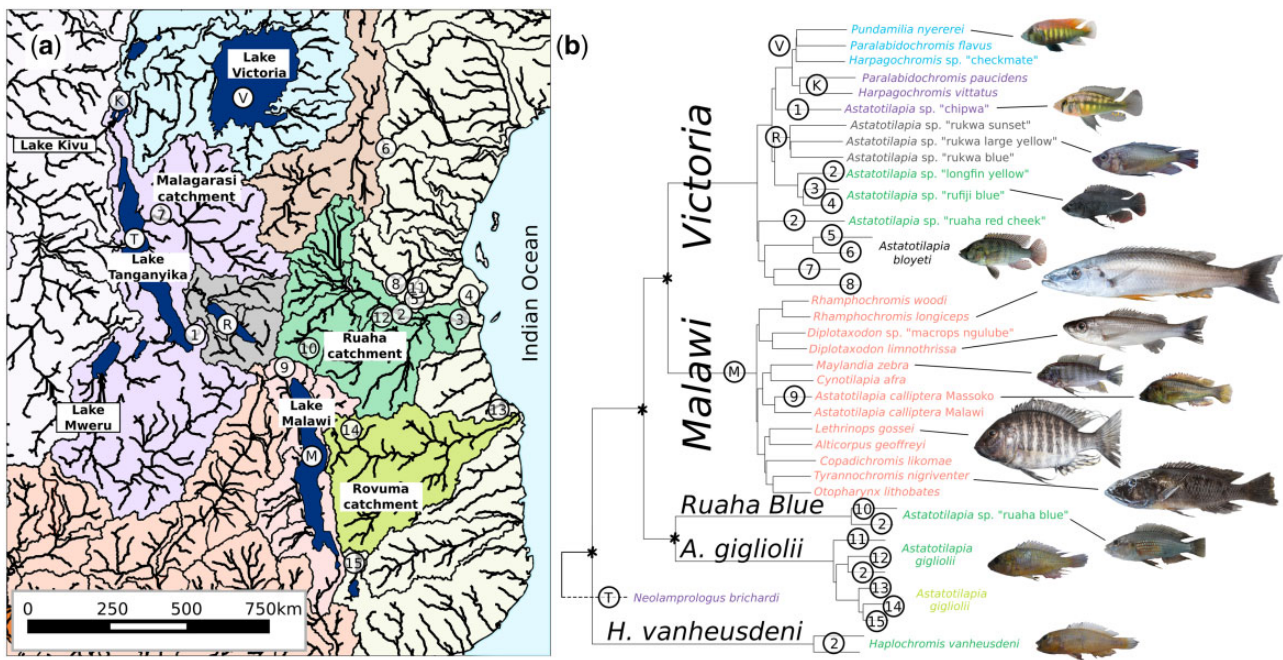
Fig. 1. Sampling and genetic relationships. (*a*) Map of East African river catchments and sampling locations. (*b*) NJ tree of pairwise genetic distances between all samples. Bootstrap support values of the nodes are given in figure 2 (for the nodes separating major clades) and in supplementary figure S1, Supplementary Material online (all nodes). The tree is rooted with the Tanganyikan outgroup *Neolamprologus brichardi*. Colors correspond to river catchments in (*a*). Black italicized clade names are used in the text to refer to the clades. The *Victoria* clade contains the LVRS and widely distributed riverine haplochromines. Some members of the *A. gigliolii* clade were previously referred to as *A. tweddlei* but after examining type specimens we suggest this to be a junior synonym of *A. gigliolii*. Relative scaling of fish pictures is to their approximate body sizes. Image credit for *Pundamilia nyererei*: O. Selz. Morte details on samples, sampling locations, and fish images can be found in supplementary table 1, Supplementary Material online.

pairwise genetic distances (fig. 1*b*). The tree suggests that the sister lineage to the Lake Malawi species flock (*Malawi* in the following) is a large clade containing the LVRS as well as widely distributed riverine haplochromines (collectively called *Victoria*, in the following). This is consistent with previous results based on AFLP and RAD sequencing data (Seehausen et al. 2003; Schwarzer et al. 2012; Meier et al. 2017). Phylogenetic splits in figure 1*b* have high bootstrap support (fig. 2*a* and supplementary fig. S1, Supplementary Material online), and the relationships between the major clades are also well supported by maximum-likelihood trees of genomic subregions (fig. 2*a*, red numbers). Furthermore, we confirmed that these species relationships also hold when using a more closely related reference genome of *Astatotilapia calliptera* (supplementary fig. S2, Supplementary Material online).

Based on mitochondrial DNA evidence, Genner et al. (2015) previously suggested that a newly discovered lineage, *Astatotilapia* sp. "ruaha blue" (*Ruaha Blue* in the following) from the Ruaha catchment, constitutes a sister lineage to the Lake Malawi radiation, rather than the above-mentioned *Victoria* clade. Although whole-genome data do not support this suggestion of a sister species relationship between the taxa (fig. 1*b*), calculation of Patterson's *D* (ABBA–BABA test) (Patterson et al. 2012) revealed a substantial excess of allele sharing between Lake Malawi samples and *Ruaha Blue* relative to *Victoria* (fig. 2*a*). Calculating the f_{branch} statistic, a summary of all possible combinations of f_4 admixture ratio

tests that are consistent with the phylogeny in figure 1*b* (Malinsky et al. 2018), revealed that this is by far the most substantial signal of excess allele sharing between any of the major taxonomic groups (supplementary fig. S3, Supplementary Material online). Using a combination of *D*-tests similar to the five-taxon test proposed by (Pease and Hahn 2015), we further confirmed that this excess allele sharing is most likely due to introgression from the *Ruaha Blue* lineage into the *Malawi* lineage (fig. 2*b*). In particular, the strongest signals correspond to excess allele sharing between *Ruaha Blue* and *Malawi* relative to *Victoria* (fig. 2*b*, red) and the second strongest to excess allele sharing between *Malawi* and *Ruaha Blue* relative to *Astatotilapia gigliolii* (fig. 2*b*, green). This suggests that the lineages leading to present-day *Ruaha Blue* and *Malawi* exchanged genes. Furthermore, we observe a significant excess of allele sharing between *A. gigliolii* and *Malawi* relative to *Victoria* (fig. 2*b*, turquoise), which is expected under gene flow from *Ruaha Blue* into *Malawi* but not for the opposite direction. Finally, the weakest signal is an excess of allele sharing between *Ruaha Blue* and *Victoria* relative to *A. gigliolii* (fig. 2*b*, blue). This signal is significant for most comparisons of samples within the clades but not for all. Although this signal cannot be explained by gene flow from *Ruaha Blue* into *Malawi*, we attribute this to further genetic exchange, for example, between riverine haplochromines in the *Victoria* clade and *Ruaha Blue*, because the signal shows clear geographic structure, with the strongest excess of allele sharing between the
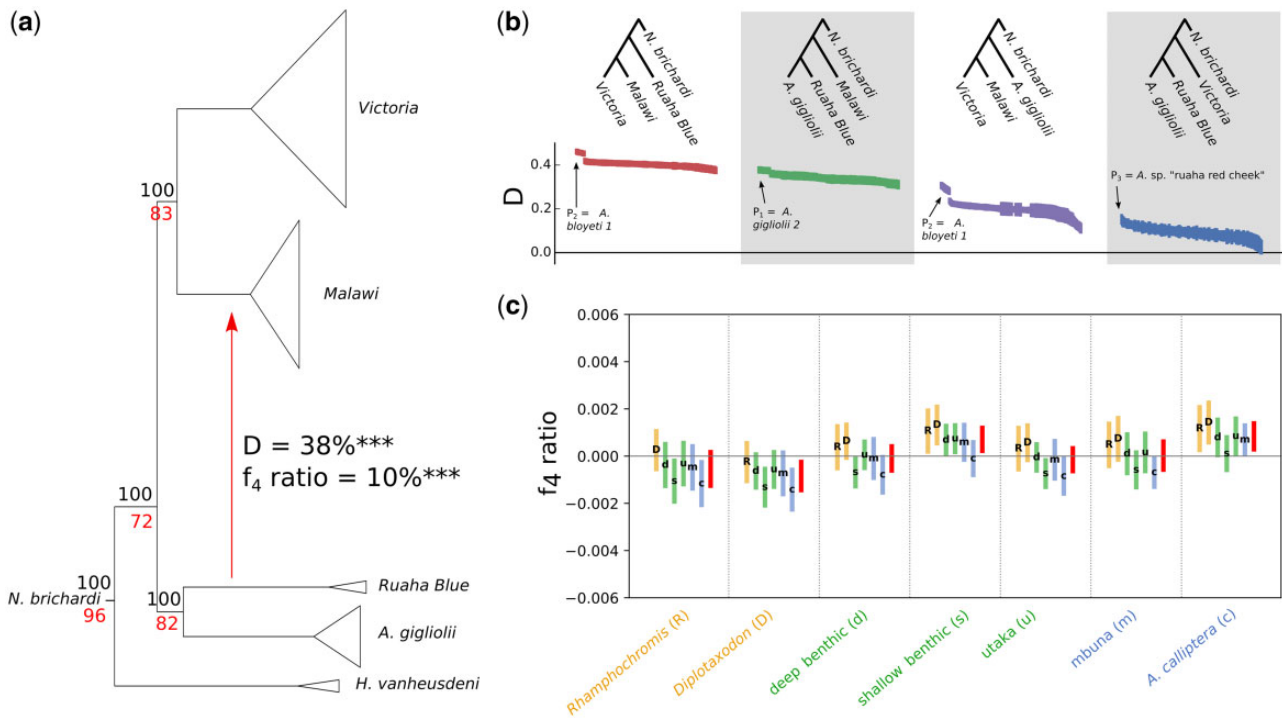
FIG. 2. Gene flow into the common ancestor of the Malawi radiation: (*a*) Cladogram of major lineages. All presented clades have 100% bootstrap support in the underlying NJ tree of all samples (black numbers above nodes; fig. 1*b*, supplementary fig. S1, Supplementary Material online; Materials and Methods). Red numbers below nodes show the percentage of times that this node was seen among 2,010 maximum likelihood gene trees of 8,000 SNP variants each (Materials and Methods). The red arrow indicates the strongest gene flow event as inferred by Patterson's D and f_4 admixture ratio ($P_1 = Victoria$, $P_2 = Malawi$; $P_3 = Ruaha\ Blue$, $P_4 = Neolamprologus\ brichardi$) (see supplementary fig. S3 and supplementary table 2, Supplementary Material online, for all f_4 admixture ratio values) (Patterson et al. 2012). ***indicates a block-jackknifing significance of $P < 10^{-16}$. (*b*) D-tests to infer the directionality of introgression. The four regions accentuated by white and gray shading correspond to $-D_{FI}$, D_{IL}, $-D_{OL}$, and D_{FO}, respectively, in the D_{FOIL} five-taxon test (*Victoria, Malawi, Ruaha Blue, A. gigliolii, N. brichardi*) (Pease and Hahn 2015), except that—as is standard for Patterson's D—only sites were considered where P_3 has the derived allele. The trees within each region show the groups from which P_1, P_2, P_3, and P_4 (from left to right) were taken for the tests below. The different data points correspond to tests for all possible combinations of samples from each of the clades. Error bars correspond to ± 3 block-jackknifing standard deviations. Outlier tests are annotated with the sample they have in common. We note that the sample *A. bloyeti* 1 from Lake Kumba in the Pangani catchment is an outlier in many comparisons consistent with it having received genetic material not related to any other sample in our data set. (*c*) f_4 admixture ratio tests (Patterson et al. 2012) with $P_3 = Victoria$ and $P_4 = Ruaha\ Blue$. P_2 is given on the *x*-axis and P_1 by the letter in the bar. Bars and labels are colored by group microhabitat (yellow = pelagic, green = benthic, blue = littoral). The red bars correspond to the case where P_1 is a group of all *Malawi* samples except P_2. Bar height corresponds to ± 3 block-jackknifing standard deviations. f_4 values between Malawi subgroups in (*c*) are a hundred times smaller than between founders in (*a*).

sympatric *Ruaha Blue* and *A.* sp. "ruaha red cheek" from the *Victoria* clade.

Using the f_4 admixture ratio (f_4 ratio) (Patterson et al. 2012), we estimated that on average ∼10% of *Malawi* genetic material traces its ancestry through the gene flow event from *Ruaha Blue* (fig. 2*a*). However, we note that in coalescent simulations this estimate is biased downwards: An f_4 admixture ratio estimate of 10% was compatible with a model with 22% actual gene flow (supplementary figs. S4 and S5, Supplementary Material online). Using a coalescent-based approach, we estimate the split time between *Victoria* and *Ruaha Blue* lineages to 3.2–3.9 Ma and the split of *Victoria* and *Malawi* to 2.0–2.8 Ma (supplementary fig. S6, Supplementary Material online).

The Lake Malawi cichlid assemblage can be organized into six genetically well-defined groups (Malinsky et al. 2018). To test whether different *Malawi* groups share different

fractions of *Ruaha Blue* alleles, we calculated f_4 ratios of the form $f_4(P_1 = Malawi-1, P_2 = Malawi-2; P_3 = Victoria, P_4 = Ruaha\ Blue)$, where *Malawi-1* and *Malawi-2* are pairs of *Malawi* groups (fig. 2*c*). Such tests are expected to be significantly positive if *Malawi-1* shares significantly more ancestry with *Ruaha Blue* than *Malawi-2* (or reciprocally negative if it shares less ancestry). These tests were mostly not significant and even those that are marginally significant are of low magnitude <0.1%, a hundred times lower than $f_4(P_1 = Victoria, P_2 = Malawi; P_3 = Ruaha\ Blue, P_4 = Outgroup)$. From this, we conclude that gene flow most likely occurred in the ancestral population prior to the radiation of the present-day *Malawi* groups and that overall there is no strong differential retention of *Victoria/Ruaha Blue*-specific alleles across *Malawi* lineages. *Victoria* and *Ruaha Blue* thus constitute extant relatives of two ancestral lineages that contributed to the Lake Malawi radiation.

Table 1. Description of the Four Variant Categories Used for the Analyses Shown in Figure 3.

No.	Name	Allelic State in				Description
		Victoria	*Malawi*	*Ruaha Blue*	*'H.' vanheusdeni*	
1	Private *Malawi*	🔵🔵	🔵🔴	🔵🔵	—	Same allele fixed in *Victoria* and *Ruaha Blue*. Expected to be enriched for young variants
			or			
		🔴🔴	🔵🔴	🔴🔴	—	
2	Shared with parent	🔵🔴	🔵🔴	—	—	Polymorphic in Malawi and at least one parental lineage. Expected to be enriched for variants that segregated in the common ancestor of *Malawi* and parental lineages
			or			
		—	🔵🔴	🔵🔵	—	
3	Hybridization-derived	🔵🔵	🔵🔴	🔴🔴	—	Differentially fixed between *Victoria* and *Ruaha Blue*. Expected to be enriched for hybridization-derived variants
			or			
		🔴🔴	🔵🔴	🔵🔵	—	
4	Differentially fixed *Victoria* –*'H.' vanheusdeni*	🔵🔵	🔵🔴	🔵🔵	🔴🔴	Same allele fixed in *Victoria* and *Ruaha Blue*; other allele fixed *'H.' vanheusdeni*. There is no evidence for hybridization of *H. vanheusdeni* with the *Malawi* ancestor. Expected to be enriched for old variation, old balanced polymorphism and technical artifacts
			or			
		🔴🔴	🔵🔴	🔴🔴	🔵🔵	

NOTE.—The first three categories are mutually exclusive and together represent all variants that are segregating in *Malawi*. The fourth category is a subset of the first. Red and blue colors correspond to two alternative alleles.

Introgressed Polymorphisms Show Excess Divergence across the Malawi Radiation

To test whether alleles that introgressed from the *Ruaha-Blue*-related lineage into the *Malawi* ancestor played a role in species divergence, we examined whether introgression-derived variants are enriched for high divergence across Malawi ecomorphological groups, adopting an approach similar to that used by Meier et al. (2017). We split variants in our data set that are polymorphic within *Malawi* into four categories depending on their allelic states and variation patterns in *Victoria*, *Ruaha Blue*, and an outgroup. The categories comprised 1) variants that are fixed for the same allele in *Victoria* and *Ruaha Blue*, 2) variants for which both *Malawi* alleles are also seen in *Victoria* and/or *Ruaha Blue*, 3) variants with the alternative alleles differentially fixed between *Victoria* and *Ruaha Blue*, and 4) variants that are fixed for the same allele in *Victoria* and *Ruaha Blue* but where the alternative allele is fixed in the outgroup '*Haplochromis*' *vanheusdeni*, respectively (table 1).

Each of these groups is expected to be enriched for different types of variants. The first and by far the largest group is enriched for young variants that arose within *Malawi* or on the ancestral branch leading to *Malawi*, whereas the second group represents shared variation that is enriched for old variants that segregated in the common ancestor of *Malawi* and *Victoria*. The third group is enriched for variants for which the polymorphism in *Malawi* was created by alleles that diverged on the branches separating the two ancestral lineages. We will therefore refer to variants in this category as "hybridization-derived." As the third pattern could in principle also be created by technical artifacts or by old polymorphism that was segregating all the way along the branch from the common ancestor of *Malawi*, *Victoria*, and *Ruaha Blue* down to *Malawi*, we added the fourth "control" category, which is expected to include both artifacts and very old polymorphisms. '*H.*' *vanheusdeni* used for this control category

does not show any signs for introgression into the *Malawi* ancestor; on the contrary, it shows significant excess allele sharing with Victoria compared with Malawi, albeit with a signal an order of magnitude smaller than the one involving *Ruaha Blue* reported above (D(*Malawi*, *Victoria*; '*H.*' *vanheusdeni*, *Neolamprologus brichardi*) = 0.04).

Next, we identified SNPs that are highly differentiated between the pelagic (*Rhamphochromis*, *Diplotaxodon*) and benthic (all other *Malawi*) lineages (*pelagic* and *benthic*, in the following) within the Lake Malawi cichlid adaptive radiation, which was suggested to be the first split in the radiation (Malinsky et al. 2018) (see also fig. 1), by taking the top 1% of SNPs ranked by fixation index (F_{ST}) value, a relative measure of genetic divergence. For each of the above-mentioned categories, we then asked what proportion of SNPs were highly differentiated (fig. 3b). We found that the set of hybridization-derived variants (category 3) is more than three times more likely to be among F_{ST} outliers compared with the set of all variants and 2.4 times more likely compared with variants shared with a parent, which represents a significant enrichment (fig. 3b; Welch's t-test $P = 10^{-13}$ and 10^{-11}, respectively). Although variants shared with a parental lineage (category 2) and variants differentially fixed between *Victoria* and '*H.*' *vanheusdeni* (category 4) also showed an excess of divergent loci compared with private *Malawi* variants, the effect was much less pronounced (1.44 and 1.16 times for categories 2 and 4, respectively, compared with 3.5 times for category 3). Similar results were obtained with different choices of F_{ST} outlier thresholds (supplementary fig. S7, Supplementary Material online), and also if the variant categories were ascertained with equally sized samples from *Victoria* and *Ruaha Blue* (supplementary fig. S8, Supplementary Material online). We also obtained similar results when we realigned a subset of the samples to a reference genome of *A. calliptera* (supplementary fig. S9, Supplementary Material online; Materials and Methods).
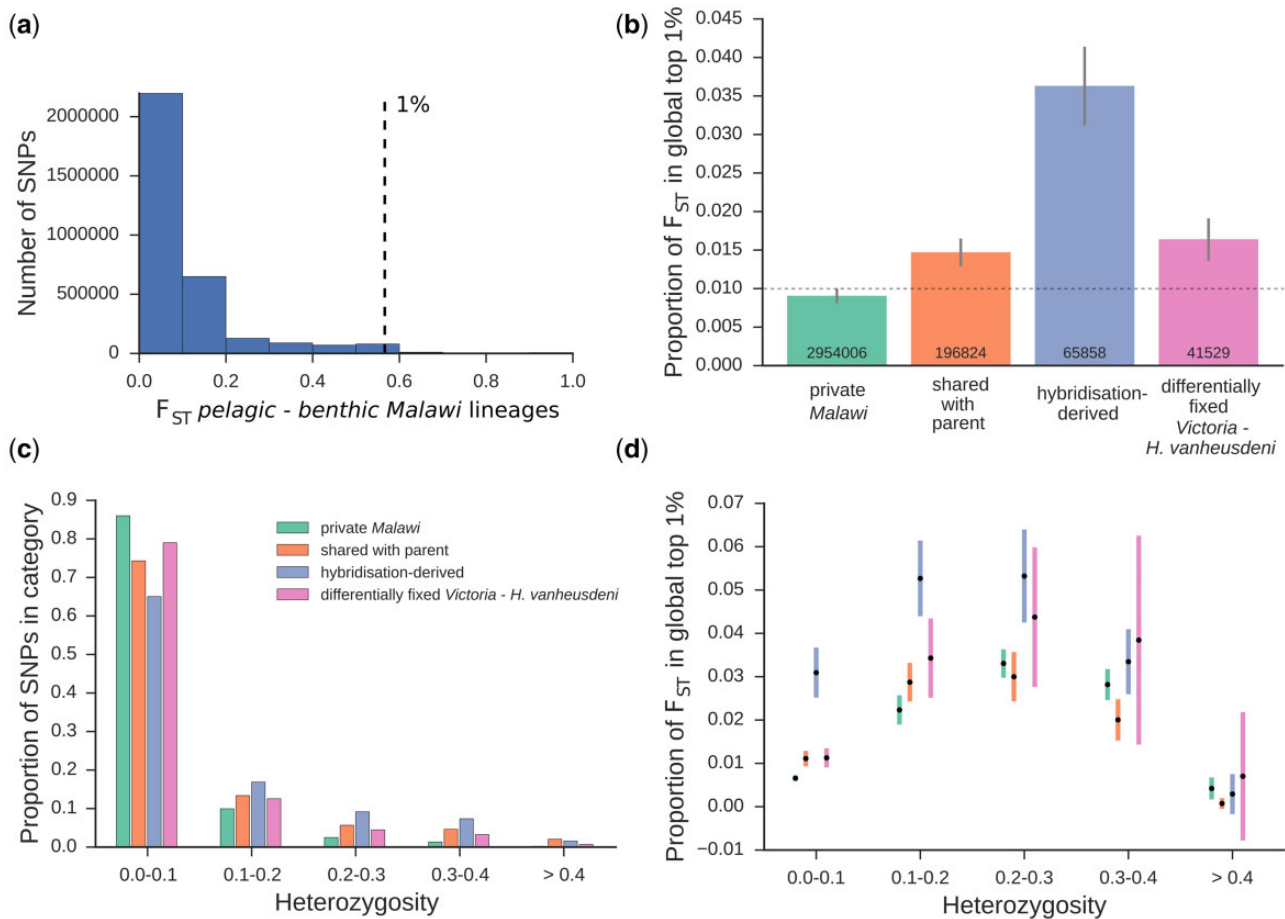
FIG. 3. Distribution of F_{ST} outliers across variant categories described in table 1. (*a*) Histogram of F_{ST} values for all SNPs; F_{ST} was calculated treating the pelagic species (genera *Rhamphochromis* and *Diplotaxodon*) and benthic species (all other *Malawi* species) as two populations, corresponding to the first split in the radiation (Malinsky et al. 2018). The dashed line shows the top 1% F_{ST} value cutoff. (*b*) For each of the variant categories described in table 1, the proportion of variants that are in the global top 1% F_{ST} outliers (right of dashed line in panel *a*) is shown. The third category is expected to be enriched for hybridization-derived variants. The numbers inside the bars correspond to the number of SNPs in each category. Gray bars correspond to ±3 block-jackknifing standard deviations. The average of the first three bars, weighted by the number of variants in the categories, is 1% by construction (dashed gray line). (*c*) For each variant category, the proportion of SNPs in different heterozygosity bins is shown. Heterozygosity is averaged across *Malawi* samples. (*d*) Proportion of F_{ST} outliers in the variant categories described in table 1 stratified by heterozygosity. Error bars correspond to ±3 block-jackknifing standard deviations.

The strong excess of highly differentiated variants among introgression-derived polymorphism suggests that these variants were under selection during the early phases of divergence. However, the excess of high F_{ST} values could also be a consequence of higher than average allele frequencies of introgression-derived variants in the ancestral population. To investigate this possibility, we assessed the distribution of average heterozygosity across *Malawi* for each of the categories, using this measure as a proxy for estimating ancestral allele frequencies (fig. 3*c*). We found that hybridization-derived variants (category 3) show elevated heterozygosity compared with the other variant categories (fig. 3*c*), as expected, but excess divergence was mainly driven by low heterozygosity variants (fig. 3*d*). Therefore, this result does not appear to be caused by elevated ancestral allele frequencies of introgression-derived variants. Further evidence for this conclusion was obtained from neutral simulations, which showed that such a disproportionate effect on divergence is not expected in the absence of selection (supplementary figs. S10–S12, Supplementary Material online).

In addition to introgression-derived variants being enriched for outlier loci, we found that these variants show elevated average F_{ST} divergence (supplementary fig. S13, Supplementary Material online) and absolute allele frequency divergence (supplementary fig. S14, Supplementary Material online). This effect is also apparent for the majority of more recent splits in *Malawi* (supplementary figs. S13 and S14, Supplementary Material online), suggesting that selection on introgression-derived variants may have played a role at multiple stages of the radiation.

Long Parental Haplotypes Are Fixed in *Malawi*— Hybridization-Derived Variants Are Interspersed along the Genome

To gain insight into the time scale of mixing of parental haplotypes in *Malawi* (i.e., *Victoria*- or *Ruaha-Blue*-related
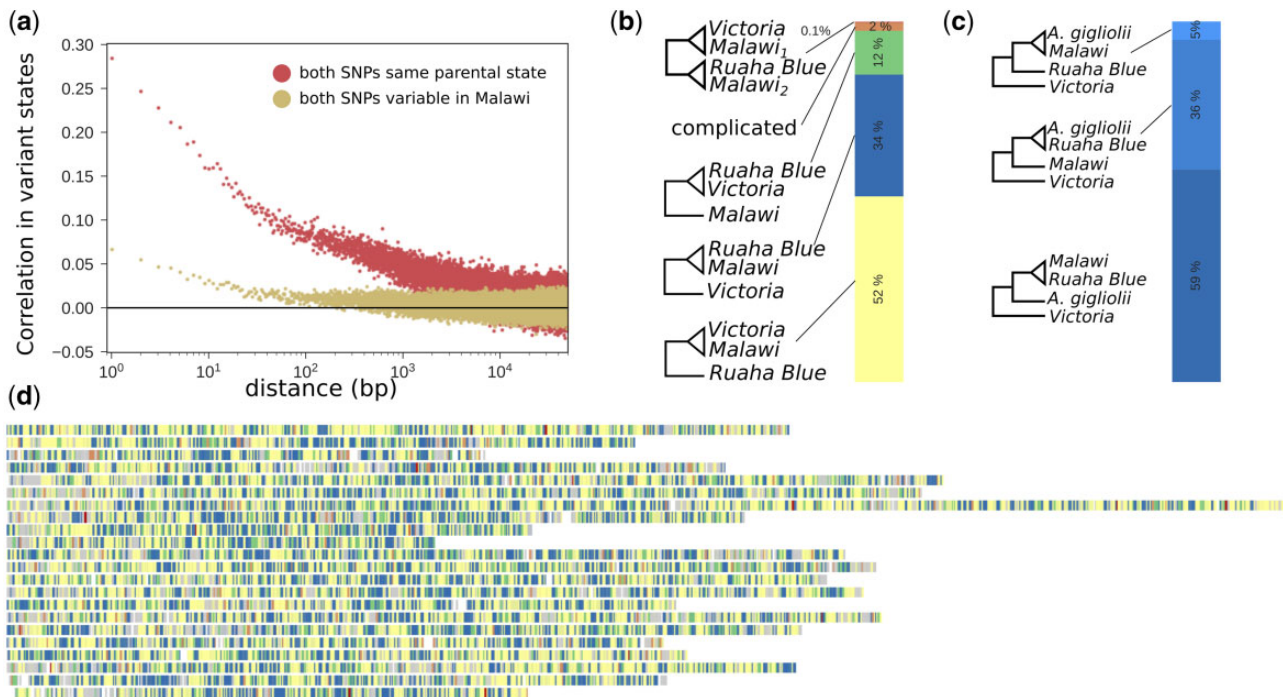
FIG. 4. Correlation in variant states and distribution of relatedness patterns along the genome. (*a*) Genome-averaged excess probability that two SNPs of a certain physical distance that are differentially fixed between *Victoria* and *Ruaha Blue* are of the same state in *Malawi*. Red dots: joint probability that two variants in *Rhamphochromis woodi* at the given distance are homozygous for the same parental state (*Victoria* or *Ruaha Blue*) minus the expectation if the two variant states were independent (Materials and Methods). Golden dots: joint probability that two variants at the given separation are variable in *Malawi* minus the expectation if the two variants were independent (Materials and Methods). (*b*) NJ trees of all samples were constructed in overlapping 5-kb windows (1 kb step size) and rooted with the Tanganyikan outgroup *Neolamprologus brichardi*. Shallow trees with <5 SNPs separating the common ancestor of *Victoria*, *Malawi*, and *Ruaha Blue* from all the tips of at least two of these groups were excluded (10% of the windows). The remaining trees were classified into five mutually exclusive patterns: 1) All *Malawi* samples are more closely related to all *Victoria* samples than to any *Ruaha Blue* sample (yellow); 2) all *Malawi* samples are more closely related to all *Ruaha Blue* samples than to any *Victoria* sample (blue); 3) all *Victoria* samples are more closely related to all *Ruaha Blue* samples than to any *Malawi* sample (green); 4) some *Malawi* samples are more closely related to all *Victoria* samples than to any *Ruaha Blue* sample, whereas other *Malawi* samples are more closely related to all *Ruaha Blue* samples than to any *Victoria* sample (red, 0.1% of the windows); and 5) more complicated relatedness patterns (brown). (*c*) Subclassification of pattern (2) into cases (from dark to light blue): All *Malawi* samples are closer to all *Ruaha Blue* samples than to any *A. gigliolii* sample; at least some *A. gigliolii* samples are closer to *Ruaha Blue*; all *Malawi* samples are closer to all *A. gigliolii* samples than to any *Ruaha Blue* sample. (*d*) Distribution of relatedness patterns along chromosomes. Regions with "shallow trees" (see above) are shown in gray.

haplotypes), we considered all fixed differences between the parental lineages, including differences not polymorphic in *Malawi*, and asked over what distance the ancestry states of SNPs in *Malawi* samples are correlated. As expected, the joint probability that two SNPs in a given *Malawi* individual are both homozygous for the *Victoria* allele or both homozygous for *Ruaha Blue* allele decreases with distance between the SNPs (fig. 4a). This correlation in ancestry states extends to several kilobases (kb) (red dots in fig. 4a), which suggests that *Malawi* individuals harbor parental haplotypes several kb in length. Recombination breaks up haplotypes at a rate of one in 30–50 megabases per generation, so this correlation pattern would correspond to haplotype fixation over a few tens of thousands of generations. In contrast to this overall pattern, fixed differences between *Victoria* and *Ruaha Blue* that are polymorphic in Malawi show much shorter range correlation in ancestry states between loci (golden dots in fig. 4a), consistent with them remaining polymorphic for much longer, allowing more time for recombination to break down linkage disequilibrium.

To better understand the genomic distribution of parental haplotypes in *Malawi*, we divided the genome into overlapping chunks of 5 kb with 4 kb overlap and constructed NJ trees of all samples for each window (fig. 4b–d). We found that in 52% of the trees all *Malawi* samples clustered with *Victoria*, in 34% of the trees all *Malawi* samples clustered with *Ruaha Blue*, in 12% of the trees *Victoria* and *Ruaha Blue* clustered together and *Malawi* samples formed an outgroup, and the remaining 2% had more complicated topologies (fig. 4b). Notably, only 0.1% of the trees showed a pattern where some *Malawi* samples clustered with *Victoria* whereas others clustered with *Ruaha Blue*, consistent with long hybridization-derived haplotypes. This is in contrast to 5.6% of the *Victoria–Ruaha Blue* fixed-differences being polymorphic in *Malawi* (i.e., in the "hybridization-derived" category). Furthermore, although the density of hybridization-derived variants is six times higher in genomic windows in which some *Malawi* samples cluster with *Victoria* and some with *Ruaha Blue* (0.82 per kb compared with 0.11 per kb and 0.15 per kb for windows in which all

Malawi samples cluster with *Victoria* and *Ruaha Blue*, respectively), more than 99% of the hybridization-derived variants are outside of these windows. These observations add further evidence that the parental haplotypes carrying hybridization-derived polymorphisms are generally much shorter than 5 kb.

Overall, there is substantial variation in the topology of relationships between *Malawi*, *Victoria*, and *Ruaha Blue* lineages at the 5-kb scale. The observation that *Malawi* clusters with *Ruaha Blue* in 34% of trees, whereas *Victoria* does so only in 12% of trees, supports a substantial contribution from the hybridization event we have described, but the presence of 12% (*Victoria*, *Ruaha Blue*) trees suggests either an additional component from incomplete lineage sorting during the separation of the three groups, or a more complex history involving multiple events. For 59% of the (*Malawi*, *Ruaha Blue*) trees, *Malawi* is closer to *Ruaha Blue* than *A. gigliolii* (fig. 4c), whereas only for 5% of these trees the opposite is true, consistent with the proposed hybridization event in the presence of some incomplete lineage sorting in the ancestry of the *Ruaha Blue*/*A. gigliolii*/*Malawi* contributor lineages. Together, this suggests that the genomes of *Malawi* samples are composed predominantly of multikilobase haplotypes deriving from *Victoria* or *Ruaha Blue* lineages. The genomic distribution of these haplotypes is nonuniform and often spans across many 5-kb windows (fig. 4d), suggesting that at least some regions fixed quite rapidly after hybridization.

Hybridization-Derived Variants Are Enriched in Vision- and Immune Defense-Related Genes

To investigate whether hybridization-derived variants contribute to specific biological functions, we tested for enrichment of hybridization-derived variants (category 3 in table 1) in genes annotated by different gene ontology (GO) terms. In particular, we computed the density of such variants across the exons of all genes (normalized by the number of private *Malawi* variants, category 1 in table 1) in each GO category and calculated an empirical enrichment *P*-value (Materials and Methods). The GO categories with the most significant enrichment are related to noncoding and ribosomal RNA functions (ribonucleoprotein complex biogenesis, $P = 0.0001$), vision (visual perception, $P = 0.0046$), and defense response (defense response to bacterium, $P = 0.005$) (fig. 5). Both genes related to vision and to pathogen defense have previously been suggested to play a role in speciation (Carleton et al. 2005; Malmstrøm et al. 2016; Egger et al. 2017; Maan et al. 2017; Meier et al. 2017; Malinsky et al. 2018). Consistent with the results above, we find that even genes in these categories do not show long hybridization-derived haplotypes, that is, most fixed differences between *Ruaha Blue* and *Victoria* in these regions are not variable in *Malawi*. However, for hybridization-derived variants there is considerable correlation of which *Malawi* individuals have *Victoria* and which have *Ruaha Blue* alleles within a gene (supplementary fig. S15, Supplementary Material online). We obtain similar enrichments when normalizing hybridization-derived variant density by the number of variants shared with a parental lineage (category 2 in table 1; supplementary

fig. S16, Supplementary Material online), or by accessible exon length (supplementary fig. S17, Supplementary Material online). In particular, in both cases visual perception is also among the most enriched GO terms ($P = 0.0006$ and 0.0101, respectively).

Hybridization-Derived Variants Assort Nonrandomly in the First Split in the Lake Malawi Cichlid Radiation

When looking at all variants, we saw in figure 2c there is not much difference in the relative amounts of *Victoria* and *Ruaha Blue* ancestries in different *Malawi* groups. Consistent with this, there is very little differential contribution of the parental lineages to *benthic* and *pelagic* (f_4(P$_1$ = *pelagic*, P$_2$ = *benthic*, P$_3$ = *Victoria*, P$_4$ = *Ruaha Blue*) = 0.08%, block-jackknifing *P*-value = 3.72×10^{-6}). However, for hybridization-derived segregating variants there is a clear excess of *Victoria* alleles in *benthic* (f_4 = 1.43%, $P = 1.42 \times 10^{-6}$). Interestingly, there is a strong positive association between allele frequency differentiation across the *benthic*/*pelagic* split and excess *Victoria* ancestry in *benthic* species (supplementary fig. S18, Supplementary Material online). For hybridization-derived variants with <10% absolute allele frequency differentiation between *benthic* and *pelagic*, there is actually a significant depletion of *Victoria* alleles in *benthic* (i.e., an excess of *Ruaha Blue*). However, this pattern reverses for more differentiated variants, with hybridization-derived variants with above 60% allele frequency difference across the first *Malawi* split showing 11% excess *Victoria* ancestry in *benthic* compared with *pelagic* (supplementary fig. S18, Supplementary Material online).

Discussion

A clear message from the increasing number of genomic investigations of nonmodel organisms is that hybridization between related species is far more common than previously thought, leading to fundamentally reticulate patterns of evolution, where species relationships are represented by networks rather than trees. In many cases, it has been shown that this process can transport selectively favorable alleles between species and thus contribute to adaptation (Herman et al. 2018; Suarez-Gonzalez et al. 2018; Walsh et al. 2018; Moest et al. 2019). However, the effect of such hybridization events on biological diversification, that is, whether it increases or reduces the rate at which new species emerge over time is not generally understood. Although empirical evidence for a direct role of hybridization in speciation events that do not involve polyploidization remains relatively rare (Schumer et al. 2014), a number of recent genomic studies suggest that hybridization of genetically divergent lineages can allow for the reassembly of old genetic variation and thereby play a decisive role in rapid adaptive diversification (reviewed in Marques et al. 2019).

Meier et al. (2017) were the first to establish a functional role of ancestral hybridization in an adaptive radiation, by showing that hybridization-derived variation was under selection during species divergence in the LVRS of cichlids. Here, we demonstrate that the Lake Malawi cichlid adaptive
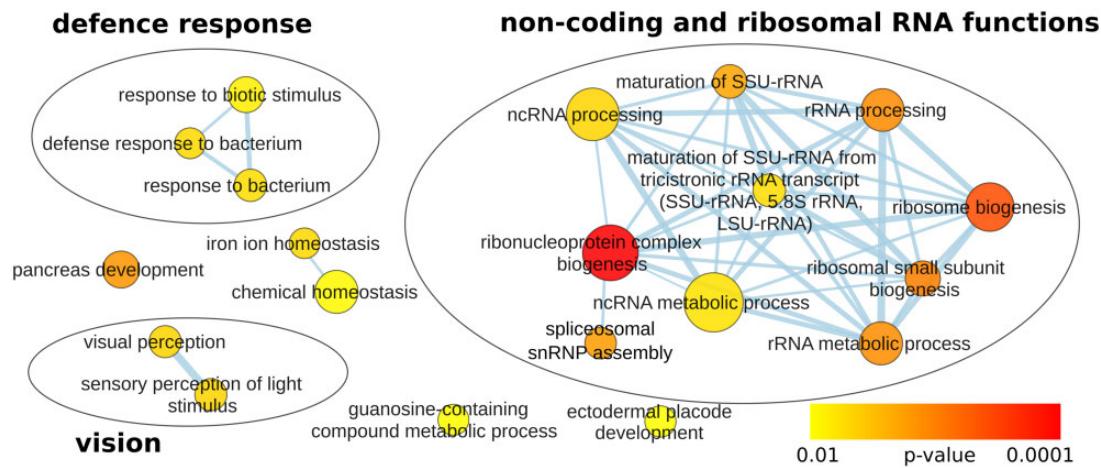
FIG. 5. Enrichment of hybridization-derived variants in biological process GOs. Genes in the *Oreochromis niloticus* gene annotation were linked to zebrafish GO categories, and enrichment of *Malawi* variants differentially fixed between *Victoria* and *Ruaha Blue* (category 3 in table 1) relative to private *Malawi* variants (category 1 in table 1) in exonic regions was tested (Materials and Methods). Categories with enrichment *P*-values <0.01 are shown. Size of the circles is proportional to the number of genes present in the annotation in each category (from 10 to 220). Analogous analyses with different normalizations are given in supplementary figures S16 and S17, Supplementary Material online.

radiation was also seeded by a hybridization event providing genetic variation that was subsequently differentially selected during divergence of the major ecomorphological clades. In particular, variants that are differentially fixed between the two parental lineages show striking excess divergence across the benthic/pelagic split of Malawi cichlids and also across more recent splits. Together with recent data from Lake Tanganyika cichlids (Irisarri et al. 2018), where hybridization at the base of radiating lineages has also been described, although not yet demonstrated to be functionally important, this suggests that hybridization might have repeatedly played an important role in the remarkable patterns of adaptive speciation in East African cichlids.

A hybrid origin of the Lake Malawi cichlid adaptive radiation had previously been proposed by Joyce et al. (2011) on the basis of mitochondrial-nuclear discordance. In particular, they suggested that two divergent riverine *A. calliptera*-like lineages seeded the species-rich Mbuna clade of Malawi cichlids. However, our recent genome-wide investigation including geographically broad sampling of *A. calliptera* has clearly shown that all *A. calliptera* lineages are nested within the Lake Malawi cichlid radiation (Malinsky et al. 2018). In contrast, the lineage we found contributing to the Malawi ancestral population—*A. sp.* "ruaha blue"—is a genome-wide outgroup to both the Lake Malawi and Lake Victoria radiations. This means that the hybridization event described here is much older than the one described by Meier et al. (2017) both in terms of the date of the event itself and in terms of the divergence between the parental lineages. Indeed, both of the parental lineages of the LVRS identified in Meier et al. (2017) are contained in our *Victoria* clade (fig. 1b, top clade).

It is notable that the present-day habitat of the only known descendant of one of the parental lineages of the Lake Malawi cichlid radiation, *A. sp.* "ruaha blue," is adjacent to that of the Lake Malawi radiation, but separated by the high-altitude Livingstone/Kipengere mountain range

estimated to have formed during the Pliocene rifting that formed Lake Malawi. However, fossil evidence supports a connection between the Ruaha/Rufiji system and the Lake Malawi basin 2–3.75 Ma (Stewart and Murray 2013), consistent with our timings of the separations of the relevant lineages.

Characterizing the distribution of parental ancestry along Malawi cichlid linkage groups, we found that the genomes of Malawi cichlids are mosaics of *Victoria*- and *Ruaha Blue*-like ancestry (the latter making up ~34%). *Ruaha Blue*-like haplotypes extend over several kbs suggesting that most polymorphism resulting from the hybridization event was fixed relatively quickly after the hybridization event in one or the other direction, within tens of thousands of years. Furthermore, there is a clear signal of nonrandom distribution of the different ancestries across the genome, which is consistent with recent observations in Princess cichlids from Lake Tanganyika (Gante et al. 2016). This might be the result of variation in local recombination rate, which has been shown to affect rates of introgression (Martin and Jiggins 2017). A high-resolution recombination map for Malawi cichlids will allow the testing of the effect of recombination rate variation on ancestry distribution in the future.

In the majority of cases, all *Malawi* samples have the same parental haplotype in a given genomic region. At first sight this seems inconsistent with the "combinatorial view on speciation and adaptive radiation" suggested by Marques et al. (2019), but it is consistent with results from the LVRS, where there is a strong correlation of parental ancestries across species (Meier et al. 2017). In Malawi cichlids, hybridization-derived segregating variants are relatively rare and interspersed among the longer, fixed parental haplotypes. A possible explanation for this pattern is that some variants managed to escape the initial fixation process through recombination or gene conversion, possibly because they were under balancing selection across a meta-population of diverging eco-types.

This would suggest that much of the hybridization-derived variation that has been maintained is functional, which could explain why, despite this ancestral hybridization event, Malawi cichlid lineages show overall remarkably little genetic diversity (Malinsky et al. 2018).

We further show that hybridization-derived variants are enriched in genes related to vision, pathogen defense, and noncoding and ribosomal RNA function. The former two categories have previously been implicated to play a role in speciation and adaptive radiation in cichlids or other fishes (Carleton et al. 2005; Malmstrøm et al. 2016; Maan et al. 2017; Meier et al. 2017; Malinsky et al. 2018). In particular, Meier et al. (2017) have shown that in the LVRS divergent parental haplotypes of a short-wavelength opsin gene differentiate clear shallow water algivores from deep turbid water detritivores. In our case, although several vision-related genes carry hybridization-derived variants, there is no separation of Malawi cichlids into species carrying Victoria-like and Ruaha Blue-like haplotypes along a whole gene body. This is consistent with our observation that hybridization-derived polymorphism is not located on long haplotypes. The observation of long-hybridization-derived haplotypes in Meier et al. (2017) is restricted to a single gene. A comparison of whole-genome sequencing data from these two independent instances of ancestral hybridization prior to radiation would allow more general understanding of the segregation patterns of parental haplotypes.

The enrichment of hybridization-derived genetic variants in noncoding RNA and ribosome-related gene ontologies was unexpected. Although we applied conservative filtering to genetic variants (see Materials and Methods), we cannot totally exclude the possibility that this pattern is related to the repetitive nature of many of the genes in these categories. However, if this pattern were due to sequencing reads from repetitive regions collapsed to the same reference position (para-SNPs), then there is no reason that this pattern would be specific to hybridization-derived variants. On the contrary, we would rather expect enrichment for genetic variants that are also polymorphic in the parental lineages. A biological explanation for the enrichment of hybridization-derived variants in noncoding-RNA-related genes might be connected to differential control of transposable elements (TEs) in the parental lineages (Wheeler 2013). Indeed, there is accumulating evidence that TEs play a role in reproductive isolation and speciation (Serrato-Capuchina and Matute 2018). This could be addressed by a thorough investigation of TE activity and distribution across Malawi cichlids and parental lineages and in crosses.

Although overall across the whole genome there is little difference in the relative contributions of parental lineages to different Malawi lineages, this is not true for hybridization-derived variants. For hybridization-derived variants that are highly differentiated between pelagic and benthic Malawi lineages (and thus expected to have been under divergent selection), it is significantly more likely to find the Victoria allele in benthic rather than in pelagic species. This signal could be driven by divergent ecological selection, for example, if the two parental lineages had different "preadaptations" to

the different niches, possibly reflecting experience of different habitats prior to introgression. This fits with the idea that introgression brings added potential by mixing up adaptive genomic combinations of the introgressing lineages (Marques et al. 2019). An alternative explanation is selection to sort out genetic incompatibilities between alleles at different loci that accumulated during divergence of the parental lineages (Stelkens et al. 2010). However, given the above observation that most parental haplotypes are actually fixed one way or the other, it is not clear whether one would expect the remaining hybridization-derived variants to show incompatibilities. We suggest that the direct investigation of present-day hybrid populations and crosses will be necessary to tease these effects apart and map genetic incompatibilities.

In this study, we infer a hybrid origin of the ancestral population of the Malawi cichlid adaptive radiation. We demonstrate that although much of the hybridization-derived variation has apparently been quickly fixed for one or the other parental haplotype in the Malawi ancestor, the variation that was maintained contributes more to early differentiation among Malawi cichlid lineages than expected in the absence of selection. Although we present the geographically most comprehensive whole-genome data set of haplochromine cichlids to date, the number of samples per species/population is mostly restricted to a single diploid individual. More comprehensive population sampling of parental lineages and species of the Lake Malawi cichlid adaptive radiation will enable more detailed characterization of hybridization-derived variants and their evolutionary significance.

Our study is based on SNPs obtained by aligning short sequencing reads to a single reference genome of the outgroup Oreochromis niloticus. The phylogenetic distance of the outgroup and the relative conservative filtering means that we are effectively investigating only around half of the genome placed on linkage groups. Although we confirm that our main results also hold when aligning to a recent Malawi cichlid reference genome (A. calliptera, Materials and Methods), for which a much larger proportion of the genome is "accessible," future investigations based on long-read sequencing will reduce the dependency on reference genomes and will also allow to investigate the effect of ancestral hybridization on structural and copy number variants.

Finally, it is worth noting that the key lineage in uncovering the evidence for introgression in the ancestor of the Malawi radiation, A. sp. "ruaha blue," appears to be represented by a single extant species confined to part of a single river system geographically remote from Lake Malawi, and entirely undocumented prior to collection of the samples presented here in 2012. Had this lineage gone extinct or not been sampled, we would have been missing direct evidence for the hybrid origin of the Malawi haplochromine lineage. This suggests that there may be ample opportunity for false negatives in the study of the influence of hybridization on adaptive radiation. Judging from our results and other recent findings in cichlid fishes and beyond, it is not impossible that hybridization indeed is the dark matter (or fuel) of rapid ecological diversification.

Materials and Methods

Sampling

Sequencing reads for some samples were obtained from previous studies (Brawand et al. 2014; McGee et al. 2016; Malinsky et al. 2018). For newly sequenced samples, ethanol-preserved fin clips were collected by M.J. Genner, A. Indermaur, A. Lamboj, B. Ngatunga, F. Ronco, W. Salzburger, and G.F. Turner between 2004 and 2014 from Malawi, Tanzania, and Zambia, in collaboration with the Fisheries Research Unit of the Government of Malawi (various collaborative projects), the Tanzania Fisheries Research Institute (MolEcoFish Project), and the Department of Fisheries, Republic of Zambia (see supplementary table 1, Supplementary Material online, for sample details and NCBI accession numbers). Samples were collected and exported with the permission of the Fisheries Research Unit of the Government of Malawi, the Tanzania Commission for Science and Technology, the Tanzania Fisheries Research Institute, and the Department of Fisheries, Republic of Zambia.

Sequencing, Variant Detection, and Filtering

New samples were sequenced on an Illumina HiSeq platform (100–125 bp paired-end reads) to a fold coverage of $\sim15\times$ (supplementary table 1, Supplementary Material online). All samples were aligned to the tilapia reference genome Orenil 1.1 (GenBank assembly accession number: GCA_000188235.2) using bwa-mem (Li 2013). Duplicate reads were marked on both per-lane and per-sample basis using the MarkDuplicates tool from the Picard software package with default options (http://broadinstitute.github.io/picard; last accessed August 2019). Local realignment around indels was performed on both per-lane and per-sample basis using the IndelRealigner tool from the GATK v3.3.0 software package. Per-sample variant detection was performed with GATK v.3.5 HaplotypeCaller (McKenna et al. 2010) and joint genotyping then performed using the GenotypeGVCFs tool. Because the data set contains samples from many species, we set a flat per-locus prior likelihoods using the –input_prior option. At this stage, we also included nonvariant sites.

Because of mapping to a relatively distant reference ($\sim3\%$ divergence in our filtered data set and $\sim6\%$ over 4-fold degenerate sites in Brawand et al. [2014]), we used more stringent filtering criteria than in previous studies (Malinsky et al. 2015, 2018). We first used Heng Li's SNPable tool (http://lh3lh3.users.sourceforge.net/snpable.shtml; last accessed August 2019), dividing the reference genome into overlapping 50-mers (subsequences of length 50) and then aligning the extracted 50-mers back to the genome (bwa aln -R 1000000 -O 3 - E 3). Then, we only kept sites in the reference for further analysis where all 50-mers mapped uniquely and without one difference hits. Next, using the all sites VCF, we masked out any sites where more than 10% of mapped reads had mapping quality zero, or where the overall mapping quality was <40, sites where the number of no-called samples was >1, sites which were within 3 bp of indels in any sample, and sites where the sum of overall depth for all samples was unusually high or low, with the depth cutoffs corresponding

to $\sim22\%$ and $\sim98\%$ percentiles of the distribution (supplementary fig. S19, Supplementary Material online). This resulted in 338 Mb of "accessible genome" on the 22 linkage groups (not including unplaced scaffolds), about half of the total placed genome size of 657 Mb. The VCF was then subset to include only SNPs, to which additional filters were applied for excess heterozygosity (InbreedingCoeff < -0.5, ExcessHet ≥ 20) and low quality by depth (DP > 2) resulting in a final data set of 18,760,777 SNPs.

To test whether aligning to a relatively distant outgroup could bias our results, we realigned a subset of the samples to a recent reference genome of the Malawi cichlid species A. calliptera (GCA_900246225.3, https://www.ncbi.nlm.nih.gov/assembly/GCF_900246225.1/; last accessed August 2019). We then called variants as described above, but used less stringent filtering criteria, only excluding sites with excessive or low coverage (site DP exceeds $\pm25\%$ of the median total site coverage) and sites with strong excess heterozygosity (ExcessHet ≥ 50). We used this callset to confirm taxon relationships (supplementary fig. S2, Supplementary Material online), gene flow events (supplementary fig. S20, Supplementary Material online), and the main result of excess divergence of hybridization-derived variants (supplementary fig. S9, Supplementary Material online). The samples used in these analyses are displayed in supplementary figure S2, Supplementary Material online.

Tree Inference, Patterson's D, and f_4 Admixture Ratio

Pairwise sequence differences were calculated and averaged over the two haplotypes of each individual using customs scripts available at https://github.com/feilchenfeldt/pypopgen; last accessed November 2019. An NJ tree was computed from the pairwise differences using the Biophyton 1.68 Phylo package. We implemented a block bootstrap by partitioning the genome into 1,000 SNP windows and calculating pairwise differences for each window. We then took 100 samples with replacement of these distance matrices, each of a size corresponding to the total number of matrices ($\sim19,000$), computed an NJ tree for each of these samples, and tested in what percentage of the samples the topology of a given node of the original tree was supported. Results are given in figure 2a and supplementary figure S1, Supplementary Material online.

RAxML trees were computed on 2,020 nonoverlapping windows of 8,000 SNPs using RAxML version 8.2.12. We then tested in what proportion of these 2,020 trees the splits between the major clades shown in figures 1b and 2a were supported (red numbers in fig. 2a).

We calculated Patterson's D and f_4 admixture ratio statistics as defined in Patterson et al. (2012) using custom scripts available at https://github.com/feilchenfeldt/pypopgen; last accessed November 2019. Calculations were performed on all SNPs that passed filtering, and block-jackknifing standard deviations were calculated across chromosomes (Green et al. 2010).

Divergence Statistics

Malawi samples were split into pelagic and benthic species, corresponding to the first split in the phylogenetic tree in

Malawi (pelagic samples: *Rhamphochromis woodi*, *R. logiceps*, *Diplotaxodon* sp. "macrops ngulube," *D. limnothrissa*). For all passed SNPs that were segregating within *Malawi*, the fixation index, F_{ST}, was calculated between pelagic and benthic species using the Weir–Cockerham estimator implemented in VCFtools 0.1.14 (Danecek et al. 2011) (option –weir-fst-pop). Absolute allele frequency differentiation of monophyletic subgroups of Malawi samples, A and B, was calculated as ABS(allele frequency A — allele frequency B) (supplementary fig. S10, Supplementary Material online).

Genomic Distribution of Parental Ancestry

We considered all variants that were differentially fixed between *Victoria* and *Ruaha Blue* (regardless of whether they were variable in *Malawi* or not) and recorded for each of these variants: 1) the state in a given focal *Malawi* individual (either homozygous for *Victoria* allele, heterozygous, or homozygous for *Ruaha Blue* allele) and (2) whether they were variable in *Malawi* or not. Then, we considered all pairs of variants of a given genomic distance in bps (only variant pairs with distance <50 kb were considered) and calculated the genome-wide fraction of variant pairs of the same state (e.g., both homozygous for the *Victoria* allele in the focal individual, both variable in *Malawi*, etc.) and the genome-wide single variant probability for each of the states. These statistics were used to calculate the empirical excess probability of observing two variants of a given distance in the same state relative to observing any two variants in the same state. In particular, the red dots in figure 4a correspond to

$$\frac{\text{\# pairs of SNPs both homozygous for } \textit{Victoria} \text{ allele}}{\text{\# of pairs}} - \left(\frac{\text{\# of SNPs homozygous for } \textit{Victoria} \text{ allele}}{\text{\# of SNPs}}\right)^2 +$$

$$\frac{\text{\# pairs of SNPs both homozygous for } \textit{RuahaBlue} \text{ allele}}{\text{\# of pairs}} - \left(\frac{\text{\# of SNPs homozygous for } \textit{RuahaBlue} \text{ allele}}{\text{\# of SNPs}}\right)^2,$$

where the focal individual was chosen as *R. woodi*. Results for different focal individuals are indistinguishable from the ones in figure 4a. Analogously, the golden dots in figure 4a correspond to

$$\frac{\text{\# pairs of SNPs both variable in } \textit{Malawi}}{\text{\# of pairs}} - \left(\frac{\text{\# of SNPs variable in } \textit{Malawi}}{\text{\# of SNPs}}\right)^2.$$

Next, we split the genome into overlapping windows of 5 kb with 1-kb offset. For each window, pairwise differences between all samples were calculated using custom scripts available at https://github.com/feilchenfeldt/pypopgen; last accessed November 2019 and NJ trees were constructed using Biopython's Phylo package. We excluded shallow trees with fewer than five SNPs separating the common ancestor of *Victoria*, *Malawi*, and *Ruaha Blue* from all the tips of at least two of these groups from further analysis (10% of the windows). This approach was chosen to remove genomic regions with shallow, uninformative trees. The remaining trees were classified according to their topology into the four patterns as described in the caption of figure 4.

Gene Enrichment Test

We downloaded the gene annotation for Orenil 1.1. release 102 from the NCBI website and retained for further analysis the 9,452 genes located on the 22 linkage groups that were annotated with an HGNC gene symbol. Using Bioconductor 3.5 in R 3.5.1, we downloaded zebrafish GO annotations (org.Dr.eg.db) for the category "biological process," compiled a mapping between GO categories and gene symbols present in Orenil 1.1 102 and retained the 1,641 GO categories with between 10 and 1,000 genes. For each gene, we computed the number of exonic variants of each of the four categories defined in table 1. Splice variants were ignored for this analysis (i.e., all exons were collapsed). We also calculated the number of accessible sites (see above) for each of the exonic regions. Next, we calculated for each GO category the total number of exonic gene variants in each of the four variant categories as well as the corresponding accessible genome lengths and computed densities of hybridization-derived variants using the following normalizations (cf., table 1): 1) (# of category 3 variants)/(# of category 1 variants), 2) (# of category 3 variants)/(# of category 2 variants), and 3) (# of category 3 variants)/(accessible genome length). To test whether these variant densities are significantly higher than expected, we computed an empirical background distribution of variant densities for randomly assembled GO categories of the same number of genes and calculated empirical P-values as $1 - \text{RANK}/10,000$, where RANK is the rank of the variant density of a GO category among the background distribution values and where 10,000 background observations were computed for each observed category size.

For each normalization, the overlap of GO categories with enrichment P-values <0.01 was visualized using the Cytoscape 3.7.1 Enrichment Map App using an edge cutoff of 0.375.

Supplementary Material

Supplementary data are available at *Molecular Biology and Evolution* online.

Acknowledgments

References

Abbott R, Albach D, Ansell S, Arntzen JW, Baird SJE, Bierne N, Boughman J, Brelsford A, Buerkle CA, Buggs R, et al. 2013. Hybridization and speciation. *J Evol Biol*. 26(2):229–246.

Barrier M, Baldwin BG, Robichaux RH, Purugganan MD. 1999. Interspecific hybrid ancestry of a plant adaptive radiation: allopolyploidy of the Hawaiian silversword alliance (Asteraceae) inferred from floral homeotic gene duplications. *Mol Biol Evol*. 16(8):1105–1113.

Berner D, Salzburger W. 2015. The genomics of organismal diversification illuminated by adaptive radiations. *Trends Genet*. 31(9):491–499.

Brawand D, Wagner CE, Li YI, Malinsky M, Keller I, Fan S, Simakov O, Ng AY, Lim ZW, Bezault E, et al. 2014. The genomic substrate for adaptive radiation in African cichlid fish. *Nature* 513(7518):375–381.

Carleton KL, Parry JWL, Bowmaker JK, Hunt DM, Seehausen O. 2005. Colour vision and speciation in Lake Victoria cichlids of the genus *Pundamilia*. *Mol Ecol*. 14(14):4341–4353.

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.

Dittrich-Reed DR, Fitzpatrick BM. 2013. Transgressive hybrids as hopeful monsters. *Evol Biol*. 40(2):310–315.

Dowling TE, DeMarais BD. 1993. Evolutionary significance of introgressive hybridization in cyprinid fishes. *Nature* 362(6419):444–446.

Edelman NB, Frandsen P, Miyagi M, Clavijo BJ, Davey J, Dikow R, Accinelli GG, Van Belleghem S, Patterson NJ, Neafsey DE, et al. 2018. Genomic architecture and introgression shape a butterfly radiation. *Science* 366(6465):594–599.

Egger B, Roesti M, Böhne A, Roth O, Salzburger W. 2017. Demography and genome divergence of lake and stream populations of an East African cichlid fish. *Mol Ecol*. 26(19):5016–5030.

Gante HF, Matschiner M, Malmstrøm M, Jakobsen KS, Jentoft S, Salzburger W. 2016. Genomics of speciation and introgression in Princess cichlid fishes from Lake Tanganyika. *Mol Ecol*. 25(24):6143–6161.

Genner MJ, Ngatunga BP, Mzighani S, Smith A, Turner GF. 2015. Geographical ancestry of Lake Malawi's cichlid fish diversity. *Biol Lett*. 11(6):20150232.

Genner MJ, Turner GF. 2012. Ancient hybridization and phenotypic novelty within Lake Malawi's cichlid fish radiation. *Mol Biol Evol*. 29(1):195–206.

Green RE, Krause J, Briggs AW, Maricic T, Stenzel U, Kircher M, Patterson N, Li H, Zhai W, Fritz M-Y, et al. 2010. A draft sequence of the Neandertal genome. *Science* 328(5979):710–722.

Heliconius Genome Consortium. 2012. Butterfly genome reveals promiscuous exchange of mimicry adaptations among species. *Nature* 487(7405):94–98.

Herman A, Brandvain Y, Weagley J, Jeffery WR, Keene AC, Kono TJY, Bilandžija H, Borowsky R, Espinasa L, O'Quin K, et al. 2018. The role of gene flow in rapid and repeated evolution of cave-related traits in Mexican tetra, *Astyanax mexicanus*. *Mol Ecol*. 27(22):4397–4416.

Irisarri I, Singh P, Koblmüller S, Torres-Dowdall J, Henning F, Franchini P, Fischer C, Lemmon AR, Lemmon EM, Thallinger GG, et al. 2018. Phylogenomics uncovers early hybridization and adaptive loci shaping the radiation of Lake Tanganyika cichlid fishes. *Nat Commun*. 9(1):3159.

Joyce DA, Lunt DH, Genner MJ, Turner GF, Bills R, Seehausen O. 2011. Repeated colonization and hybridization in Lake Malawi cichlids. *Curr Biol*. 21(3):R108–R109.

Kocher TD. 2004. Adaptive evolution and explosive speciation: the cichlid fish model. *Nat Rev Genet*. 5(4):288–298.

Kozak KM, McMillan O, Joron M, Jiggins CD. 2018. Genome-wide admixture is common across the *Heliconius* radiation. doi: 10.1101/414201. http://dx.doi.org/10.1101/414201.

Lamichhaney S, Berglund J, Almén MS, Maqbool K, Grabherr M, Martinez-Barrio A, Promerová M, Rubin C-J, Wang C, Zamani N, et al. 2015. Evolution of Darwin's finches and their beaks revealed by genome sequencing. *Nature* 518(7539):371–375.

Li H. 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv [q-bioGN]. Available from: http://arxiv.org/abs/1303.3997.

Maan ME, Seehausen O, Groothuis T. 2017. Differential survival between visual environments supports a role of divergent sensory drive in cichlid fish speciation. *Am Nat*. 189(1):78–85.

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.

Malinsky M, Svardal H, Tyers AM, Miska EA, Genner MJ, Turner GF, Durbin R. 2018. Whole-genome sequences of Malawi cichlids reveal multiple radiations interconnected by gene flow. *Nat Ecol Evol*. 2(12):1940.

Mallet J. 2007. Hybrid speciation. *Nature* 446(7133):279–283.

Mallet J, Besansky N, Hahn MW. 2016. How reticulated are species? *Bioessays* 38(2):140–149.

Malmstrøm M, Matschiner M, Tørresen OK, Star B, Snipen LG, Hansen TF, Baalsrud HT, Nederbragt AJ, Hanel R, Salzburger W, et al. 2016. Evolution of the immune system influences speciation rates in teleost fishes. *Nat Genet*. 48(10):1204–1210.

Marques DA, Meier JI, Seehausen O. 2019. A combinatorial view on speciation and adaptive radiation. *Trends Ecol Evol*. 34(6):531–544.

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

McGee MD, Neches RY, Seehausen O. 2016. Evaluating genomic divergence and parallelism in replicate ecomorphs from young and old cichlid adaptive radiations. *Mol Ecol*. 25(1):260–268.

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, et al. 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res*. 20(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(1):14363.

Meyer BS, Matschiner M, Salzburger W. 2016. Disentangling incomplete lineage sorting and introgression to refine species-tree estimates for Lake Tanganyika cichlid fishes. *Syst Biol*. 66(4):531–550.

Moest M, Van Belleghem SM, James JE, Salazar C, Martin SH, Barker SL, Moreira GRP, Mérot C, Joron M, Nadeau NJ, et al. 2019. Classic and introgressed selective sweeps shape mimicry loci across a butterfly adaptive radiation. doi: 10.1101/685685. http://dx.doi.org/10.1101/685685.

Novikova PY, Hohmann N, Nizhynska V, Tsuchimatsu T, Ali J, Muir G, Guggisberg A, Paape T, Schmid K, Fedorenko OM, et al. 2016. Sequencing of the genus *Arabidopsis* identifies a complex history of nonbifurcating speciation and abundant trans-specific polymorphism. *Nat Genet*. 48(9):1077–1082.

Patterson N, Moorjani P, Luo Y, Mallick S, Rohland N, Zhan Y, Genschoreck T, Webster T, Reich D. 2012. Ancient admixture in human history. *Genetics* 192(3):1065–1093.

Pease JB, Hahn MW. 2015. Detection and polarization of introgression in a five-taxon phylogeny. *Syst Biol*. 64(4):651–662.

Richards EJ, Martin CH. 2017. Adaptive introgression from distant Caribbean islands contributed to the diversification of a microendemic adaptive radiation of trophic specialist pupfishes. *PLoS Genet*. 13(8):e1006919.

Rieseberg LH, Archer MA, Wayne RK. 1999. Transgressive segregation, adaptation and speciation. *Heredity* 83(4):363–372.

Salzburger W. 2018. Understanding explosive diversification through cichlid fish genomics. *Nat Rev Genet*. 19(11):705–717.

Schumer M, Cui R, Rosenthal GG, Andolfatto P. 2015. Reproductive isolation of hybrid populations driven by genetic incompatibilities. *PLoS Genet*. 11(3):e1005041.

Schumer M, Rosenthal GG, Andolfatto P. 2014. How common is homoploid hybrid speciation? *Evolution* 68(6):1553–1560.

Schwarzer J, Swartz ER, Vreven E, Snoeks J, Cotterill FPD, Misof B, Schliewen UK. 2012. Repeated trans-watershed hybridization among haplochromine cichlids (Cichlidae) was triggered by Neogene landscape evolution. *Proc R Soc B.* 279(1746):4389–4398.

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

Seehausen O. 2013. Conditions when hybridization might predispose populations for adaptive radiation. *J Evol Biol.* 26(2):279–281.

Seehausen O. 2015. Process and pattern in cichlid radiations—inferences for understanding unusually high rates of evolutionary diversification. *New Phytol.* 207(2):304–312.

Seehausen O, Koetsier E, Schneider MV, Chapman LJ, Chapman CA, Knight ME, Turner GF, van Alphen JJM, Bills R. 2003. Nuclear markers reveal unexpected genetic variation and a Congolese-Nilotic origin of the Lake Victoria cichlid species flock. *Proc R Soc Lond B.* 270(1511):129–137.

Serrato-Capuchina A, Matute DR. 2018. The role of transposable elements in speciation. *Genes* 9(5):254.

Stelkens RB, Young KA, Seehausen O. 2010. The accumulation of reproductive incompatibilities in African cichlid fish. *Evolution* 64(3):617–633.

Stewart KM, Murray AM. 2013. Earliest fish remains from the Lake Malawi Basin, Malawi, and biogeographical implications. *J Vertebr Paleontol.* 33(3):532–539.

Stryjewski KF, Sorenson MD. 2017. Mosaic genome evolution in a recent and rapid avian radiation. *Nat Ecol Evol.* 1(12):1912–1922.

Suarez-Gonzalez A, Lexer C, Cronk Q. 2018. Adaptive introgression: a plant perspective. *Biol Lett.* 14(3):20170688.

Svardal H, Jasinska AJ, Apetrei C, Coppola G, Huang Y, Schmitt CA, Jacquelin B, Ramensky V, Müller-Trutwin M, Antonio M, et al. 2017. Ancient hybridization and strong adaptation to viruses across African vervet monkey populations. *Nat Genet.* 49(12):1705–1713.

Wagner CE, Harmon LJ, Seehausen O. 2012. Ecological opportunity and sexual selection together predict adaptive radiation. *Nature* 487(7407):366–369.

Walsh J, Kovach AI, Olsen BJ, Shriver WG, Lovette IJ. 2018. Bidirectional adaptive introgression between two ecologically divergent sparrow species. *Evolution* 72(10):2076–2089.

Wheeler BS. 2013. Small RNAs, big impact: small RNA pathways in transposon control and their effect on the host stress response. *Chromosome Res.* 21(6–7):587–600.