Different Sources of Allelic Variation Drove Repeated Color Pattern Divergence in Cichlid Fishes

Sabine Urban, Alexander Nater, Axel Meyer, and Claudius F. Kratochwil

1Department of Biology, University of Konstanz, Konstanz, Germany

*Corresponding authors: E-mails: claudius.kratochwil@gmail.com; axel.meyer@uni-konstanz.de.

Associate editor Yoko Satta

Abstract

The adaptive radiations of East African cichlid fish in the Great Lakes Victoria, Malawi, and Tanganyika are well known for their diversity and repeatedly evolved phenotypes. Convergent evolution of melanic horizontal stripes has been linked to a single locus harboring the gene agouti-related peptide 2 (agrp2). However, where and when the causal variants underlying this trait evolved and how they drove phenotypic divergence remained unknown. To test the alternative hypotheses of standing genetic variation versus de novo mutations (independently originating in each radiation), we searched for shared signals of genomic divergence at the agrp2 locus. Although we discovered similar signatures of differentiation at the locus level, the haplotypes associated with stripe patterns are surprisingly different. In Lake Malawi, the highest associated alleles are located within and close to the 5′ untranslated region of agrp2 and likely evolved through recent de novo mutations. In the younger Lake Victoria radiation, stripes are associated with two intronic regions overlapping with a previously reported cis-regulatory interval. The origin of these segregating haplotypes pre-dates the Lake Victoria radiation because they are also found in more basal riverine and Lake Kivu species. This suggests that both segregating haplotypes were present as standing genetic variation at the onset of the Lake Victoria adaptive radiation with its more than 500 species and drove phenotypic divergence within the species flock. Therefore, both new (Lake Malawi) and ancient (Lake Victoria) allelic variation at the same locus fueled rapid and convergent phenotypic evolution.

Key words: standing genetic variation, cichlid fishes, convergent evolution, color patterns, evolutionary genomics, adaptive radiations.

Introduction

Understanding how genetic variation translates into phenotypic diversity is an important goal in evolutionary biology. Repeatedly evolved phenotypes are particularly interesting for the study of the genetic basis of phenotypic diversity because they provide natural replicates that can inform whether the same evolutionary mechanisms have recurrently generated these phenotypes (Kuraku and Meyer 2008; Protas and Patel 2008; Stern 2013; Elmer and Meyer 2011; Elmer et al. 2014; Kratochwil and Meyer 2015; Kratochwil et al. 2018). Repeated evolution can result either from evolution through independent de novo mutations occurring in different species; or from preexisting variation that can be recruited via introgression or from standing genetic variation in a common ancestor (Stern 2013). Most de novo mutations are expected to be neutral or deleterious (Ohta 1992), whereas old standing genetic variation has likely already been purged from deleterious alleles due to previous selection. Adaptation from standing genetic variation is generally thought to be faster, as alleles reach fixation more quickly. Thereby, standing genetic variation might facilitate rapid diversification (Barrett and Schluter 2008; Marques et al. 2019). Ancestral variants can be more easily reassembled into new combinations, whereas fixation of de novo mutations is predicted to result in a slower diversification process (Barrett and Schluter 2008; Hedrick 2013; Marques et al. 2019). Accordingly, recent studies recognized the recruitment of alleles from standing genetic variation as an important evolutionary mechanism in driving the rapid phenotypic diversification found in adaptive radiations (Colosimo et al. 2004; Hines et al. 2011; Seehausen 2015; Lamichhaney et al. 2016; Han et al. 2017; Meier, Marques, et al. 2017; Bassham et al. 2018; Malinsky et al. 2018; Nelson and Cresko 2018; Salzburger 2018; York et al. 2018; Lewis et al. 2019; Svardal et al. 2020; Kau et al., forthcoming). However, with a few exceptions (Colosimo et al. 2004; Hines et al. 2011; Lamichhaney et al. 2016; Meier, Marques, et al. 2017; Lewis et al. 2019), most studies reporting evidence for the importance of ancestral standing genetic variation across whole genomes lacked knowledge of genotype–phenotype connections (Malinsky et al. 2018; Nelson and Cresko 2018; Svardal et al. 2020). As a consequence, the specific impact of old genetic variation on phenotypic diversification often remains elusive.

The adaptive radiations of cichlid fishes offer a great opportunity to investigate the contribution of standing genetic variation to rapid adaptive divergence, due to their unusually high diversity in species and the repeated evolution of
multiple phenotypes (Meyer et al. 1990; Meyer 1993; Stiassny and Meyer 1999; Kocher 2004; Genner and Turner 2005; Henning and Meyer 2014). Within the Great Lakes of the African Rift Valley, cichlids diversified into hundreds of endemic species in several lakes of different sizes and ages. In the three East African Great Lakes alone—Lake Victoria, Lake Tanganyika, and Lake Malawi—more than 1,200 cichlid species evolved (Salzburger and Meyer 2004). Recent studies showed that the onset of the exceptionally rapid adaptive radiation in Lake Victoria, in which at least 500 species evolved within the past 15,000 years (Johnson et al. 2000; Verheyen et al. 2003; Wagner et al. 2013), was fueled by high levels of genome-wide standing genetic variation (Seehausen 2004; Meier, Marques, et al. 2017). The Lake Victoria cichlid flock is derived from divergent lineages of the geologically older Lake Kivu (Verheyen et al. 2003) and adjacent rivers (Salzburger et al. 2005; Meier, Marques, et al. 2017), which started diversifying about 100–200 ky ago (Verheyen et al. 2003; Seehausen 2006; Genner et al. 2007). The older radiation of Lake Malawi cichlids encompasses about 700 species (Turner et al. 2008), which are believed to have evolved within the last 800 ky (Meyer et al. 1990; Danley and Kocher 2001; Brawand et al. 2014). Recently, whole-genome resequencing revealed that standing genetic variation contributed to the high diversification rates of this adaptive radiation (Svardal et al. 2020). Furthermore, standing genetic variation derived from ancestral lineages was also reported for Lake Tanganyika cichlids (Irisarri et al. 2018), the oldest and phenotypically most diverse of the three East African cichlid fish adaptive radiations (Sturmbauer and Meyer 1992; Salzburger et al. 2005; Koblmüller et al. 2008).

Within and between these different adaptive radiations multiple phenotypes have evolved repeatedly (Stiassny and Meyer 1999). This is exemplified by melanic horizontal stripes, an adaptive phenotype that is often associated with shoaling behavior and a piscivorous feeding mode (Seehausen and Alphen 2001). Previous work identified the gene agouti-related peptide 2 (agrp2, also called asip2b) as a major effect locus for stripe pattern divergence in African cichlids (Kratochwil et al. 2018). The teleost-specific gene agrp2/asip2b and its paralogs have been previously associated with pigmentation phenotypes (Zhang et al. 2010; Manceau et al. 2011; Ceinos et al. 2015). In zebrafish, agrp2 is mainly expressed in the pineal gland. Biochemically it acts as an antagonist of melanocortin receptors (Zhang et al. 2010). In cichlids, agrp2 has been demonstrated to also have a function in the skin, where it controls the presence of stripe patterns. High expression of agrp2 inhibits stripe patterns, while low expression permits their development. Yet, prior work could not identify the exact causal haplotypes and their evolutionary origin(s). The adaptive importance of horizontal stripes (Seehausen and Alphen 2001), together with a well-resolved genotype-to-phenotype connection (Kratochwil et al. 2018), makes the agrp2 locus an ideal target to investigate the role of preexisting standing genetic variation versus de novo mutations for driving adaptive phenotypic divergence in rapidly evolving species flocks.

Here, we addressed whether striped and nonstriped fish of the parallel and independent radiations of Lake Victoria, Lake Malawi, and Lake Tanganyika show the same signals of genomic divergence explaining the convergent evolution of stripe patterns. By including genomic sequences of more basal lineages of the Lake Victoria radiation from Lake Kivu and adjacent rivers we furthermore traced back the evolutionary origin of the causal major effect allelic variants of Lake Victoria cichlids.

**Results and Discussion**

*Stripe Pattern Convergence and Diversification in African Cichlid Fish Radiations*

To reconstruct the evolutionary history of the haplotypes associated with stripe patterns, we investigated the genomic interval around the agrp2 gene with a combination of target enrichment (~30-kb agrp2 region ± 100 kb) and whole-genome resequencing as the agrp2 locus was previously shown to be associated with horizontal stripes in cichlids of the three African Great Lakes (Henning et al. 2014; Kratochwil et al. 2018). Data were collected from 213 individuals from the three great African species flocks (number of individuals/species in Lake Malawi n = 143/111, Lake Tanganyika n = 26/23, Lake Victoria n = 36/22; supplementary table S1, Supplementary Material online).

We inferred a species tree of the sampled species based on 6,545 genome-wide randomly selected loci of 3 kb from 33 high-quality genomes. The phylogeny from this state-of-the-art, high-density data set agrees with previous reports based on mitochondrial (Meyer et al. 1990), RAD-seq (Wagner et al. 2013) and, most recently, genomic data (Malinsky et al. 2018; Svardal et al. 2020). All phylogenies show strong discordance between the stripe phenotype and phylogeny showing that stripes clearly evolved repeatedly (fig. 1).

*Stripes in Lake Malawi and Victoria Radiations Are Associated with the Same Gene but Different Noncoding Regions*

Using whole-genome resequencing and target enrichment data, we calculated relative genetic differentiation ($F_{ST}$) between striped and nonstriped species for the cichlid radiations of Lakes Tanganyika, Malawi, and Victoria over the 672,091 filtered biallelic single-nucleotide polymorphisms (SNPs) called across the whole ~10-Mb scaffold 3 containing the agrp2 gene. Parallel evolution drives certain mutations to fixation independently in different populations and thereby acts on very local genomic regions. Therefore, we used the software Sauguro for $F_{ST}$ calculation which implements an algorithm that sets out to identify and pinpoint such regions using a Hidden Markov Model and a Neural Network, applied in an interleaved fashion. Sauguro then infers local relationships among individuals in the form of genetic distance matrices and assigns segments across the genomes to these topologies.

In the Lake Tanganyika radiations, we did not find regions of elevated $F_{ST}$ between striped and nonstriped species around agrp2 (supplementary fig. S1, Supplementary Material online), although a link between agrp2 expression and stripes has been shown earlier (Kratochwil et al. 2018).
The Lake Tanganyika species flock is more than 10 My old and consists of several ancient independent radiations (Salzburger et al. 2002, 2005; Clabaut et al. 2005; Koblmüller et al. 2008; Takahashi and Koblmüller 2011) with a complex history of repeated colonization events (Nishida 1991; Salzburger et al. 2002). Therefore, the missing association of alleles within the agrp2 locus with stripes is likely explained by more complex genetic mechanisms of stripe formation and might involve multiple cis-regulatory loci and/or trans-regulatory mechanisms as well as additional modifier loci.

Both adaptive radiations of Lakes Victoria and Malawi are composed of a single lineage of haplochromine cichlids which evolved within the last 2–4 My in Lake Malawi and in Lake Victoria within 0.01–1 My (Meyer et al. 1990; Kocher 2004; Turner 2007; Brawand et al. 2014). Among the 700 endemic Lake Malawi cichlids, the agrp2 locus shows elevated differentiation among the littoral rock-dwelling mbuna which contains at least 200 species (fig. 2A) (Danley and Kocher 2001). Within this lineage, the strongest differentiation between striped and nonstriped species includes the 5′ untranslated region (UTR) of agrp2 ($F_{ST} = 0.85$ vs. scaffold mean 0.09; fig. 2A and B). In a gene tree inferred based on this region, the topology clearly separates striped mbuna from nonstriped mbuna but not the Lake Victoria phenotypes (fig. 2C). A single species, Petrotilapia nigra, is heterozygous for two of the three variants close to and within the 5′-UTR (supplementary fig. S2, Supplementary Material online) and shows a very indistinct stripe pattern. Variants within region LM are unique to striped Lake Malawi mbuna, and there is no association in nonmbuna or Lake Victoria cichlids (fig. 3). These variants, therefore, most likely constitute de novo mutations that evolved within the last 300 ky in the Lake Malawi mbuna radiation (Genner et al. 2007). However, this association between the agrp2 locus and stripes vanishes when comparing the whole Lake Malawi data set including nonmbuna species (supplementary fig. S1, Supplementary Material online).

For the Lake Victoria radiation, the agrp2 locus was shown to be highly differentiated between striped and nonstriped species (fig. 2A). The two most differentiated regions ($F_{ST} = 0.87$ and $F_{ST} = 0.78$ vs. scaffold mean of 0.06) are directly upstream of the second exon and largely overlap (58.4% overlap) with a cis-regulatory active region (442,318–443,409) that was previously identified based on Sanger sequencing of three Lake Victoria species and experimentally tested using a transgenic reporter assay (Kratochwil et al. 2018). Taken together, the Lake Victoria regulatory interval (including both highly associated regions, LV 1 and LV 2; fig. 2B) has a size of ~1.23 kb and is likely composed of several smaller cis-regulatory elements such as enhancers and/or silencers (fig. 2B). In contrast to the topology of the region LM, the gene trees inferred from these two regions of highest differentiation (LV 1 and LV 2) clearly separate striped from nonstriped
The same gene but different lake-specific regulatory regions are associated with the repeated evolution of stripe patterns in the adaptive radiations of haplochromine cichlids of Lakes Malawi and Victoria. (A) Association of stripes with genomic regions. Black dots represent midpoints of every associated region ($F_{ST}$ value) as identified by *Saguaro* and black lines are smoothened local regressions between striped and nonstriped species from Lake Victoria (top) and Lake Malawi mbuna species (bottom). This is plotted together with topology weights for topologies in which striped and nonstriped species are reciprocally monophyletic (orange bars Lake Victoria, blue bars Lake Malawi mbuna). Each value gives the proportionate contribution of a particular taxon tree to the full tree with values ranging from 0 to 1. An example for such a topology in which striped species are reciprocally monophyletic is provided for both radiations. (B) Gene structure of *agrp2* with regions of elevated $F_{ST}$ ($F_{ST} > 0.75$).
Lake Victoria cichlids (fig. 2D and E). This pattern supports the hypothesis that different regulatory regions at the same locus facilitate convergent evolution of stripe patterns across different cichlid radiations.

To further support the association of the identified cis-regulatory intervals found in Lakes Malawi and Victoria with stripe patterns, we employed a second, complementary approach, in which we assessed topology weights with TWISST (Van Belleghem et al. 2017). The TWISST results strongly support a topology that groups species by stripe phenotype (fig. 2A). The adjacent gene, *atp6V0d2*, also exhibited pronounced topology grouping by stripe phenotype, but previous work did not reveal any fixed mis- or nonsense mutations or differential expression (Kratochwil et al. 2018).

Noncoding Variants Predict Changes in Transcription Factor Binding in Highly Divergent Regions of Both, Lakes Malawi and Victoria

Both highly divergent regions are noncoding and might therefore contribute to variation in *agrp2* transcription and/or translation. The highly associated 90-bp region in Lake Malawi (LM; fig. 2B and C) overlaps with the 5′-UTR of *agrp2*. 5′-UTRs can contain transcription factor-binding sites (TFBSs) (Barrett et al. 2012; Lavallee-Adam et al. 2017) but also have been shown to play important roles in posttranscriptional regulation (Araujo et al. 2012) and could therefore lead to variation in transcript stability or translation rate. To provide additional evidence that the substitutions within and close to the 5′-UTR of *agrp2* in Lake Malawi mbuna might influence *agrp2* transcription, we screened these regions for potential TFBSs that are likely affected by the associated variants. For this, we used sequences from a representative non-striped and striped species (nonstriped *Pseudotropheus demasoni*, and striped *Ps. cyanorhabdos*) flanking the three variant sites within candidate region LM ±10 bp (fig. 3 and supplementary fig. S2 and table S2, Supplementary Material online). The sequence flanking the first variant (position 438,598) contained 18 TFBSs of which ten have a delta relative score of >0.1 (Materials and Methods, supplementary table S2 and fig. S2, Supplementary Material online) therefore suggesting a higher TF-binding affinity in the nonstriped (the species with high expression of the “stripe-repressor gene” *agrp2*) than in the striped species. The TFs include *tf3* that was associated with pigmentation previously (Dorsky et al. 2012; Noor et al. 2016; Lavallee-Adam et al. 2017).

**Fig. 3.** Different lake-specific substitutions are associated with stripe pattern evolution in Lakes Malawi and Victoria. (A) Gray represents the “nonstriped” allele of the *Pundamilia nyererei* reference genome. All nonreference alleles within candidate region LM are indicated in blue. Substitutions in LM are not shared between Lake Victoria and Lake Malawi cichlids suggesting that they originated de novo in Lake Malawi mbuna. Orange indicates substitutions that were found in candidate region LV 1 that are associated with stripe patterns in Lake Victoria cichlids. Although those haplotypes are divergent between striped and nonstriped cichlids in Lake Victoria, this is not the case in Lake Malawi mbuna. Black arrows highlight substitutions that are highly divergent between striped and nonstriped species (>90% frequency difference since some striped species are heterozygous, see also fig. 4A and supplementary fig. S3, Supplementary Material online). (B) A screen for potential TFBSs in a nonstriped (*Pseudotropheus demasoni*, *Pdem*, in Malawi and *Pundamilia nyererei*, *Pnye*, in Victoria) and striped (*Pseudotropheus cyaneorhabdos*, *Pcy*, in Malawi and *Haplochromis sauvagei*, *Hsau*, in Victoria) representative species reveals several TFBSs that are affected by stripe-associated alleles. Shown are those TFs with the highest delta relative score. Variable positions in the motif sequence are capitalized and underlined.
whereas transregulatory factors might be identical in both radiations nfix, spi1, nr2c2(var.2), zeb1, rbpj, sox3, sox10. TF-binding differences are shared between LM and LV (i.e., melanocyte/melanophore life cycle (Harris et al. 2010)). Some including Sox18 regulate and interact during all stages of the family of TFs with lower delta relative scores. Sox proteins position 442,399 contains six more TFBSs belonging to the sox position 442,399 has a TFBS for desmelto th es a m eg e n ed o e sn o tn e c e s s a r i lyme a nt h a ta l s o expression, demonstrating that the recurrent involve- position 442,399 is LV 1. We therefore also screened this region for potential TFBSs using the nonstriped species Pundamilia nyererei (Pnye) and striped Haplochromis sauvagei (Hsau). Our analysis revealed high delta relative scores for several TFs with a known function in pigmentation pathways. For example, tcfl5 at position 441,862 belongs to a group of TFs involved in the Wnt signaling pathway. In zebrafish, Wnt signaling activates nacre, a zebrafish homolog of mitf, a key regulator of pigment synthesis, which in turn leads to pigment cell differentiation. Position 442,188 harbors a TFBS for zeb1, which in cichlids represses the expression of mitf (Albertson et al. 2014). The sequence around variant position 442,399 has a TFBS for sox18. The sequence around position 442,399 contains six more TFBSs belonging to the sox family of TFs with lower delta relative scores. Sox proteins including Sox18 regulate and interact during all stages of the melanocyte/melanophore life cycle (Harris et al. 2010). Some TF-binding differences are shared between LM and LV (i.e., nfix, spi1, nr2c2(var.2), zeb1, rbpj, sox3, sox10) suggesting that transregulatory factors might be identical in both radiations whereas cis-regulatory elements are not.

These analyses support that divergence between striped and nonstriped species in the radiations of East African cichlids is fueled by distinct cis-regulatory mechanisms controlling agrp2 expression, demonstrating that the recurrent involvement of the same gene does not necessarily mean that also the underlying causal mutations are the same.

The Causal Stripe Haplotype in Lake Victoria Evolved Prior to the Adaptive Radiation
The finding of a single haplotype associated with stripes across all Lake Victoria species in our data set is particularly interesting, as it suggests recruitment from ancestral standing genetic variation that was already present prior to the Lake Victoria cichlid radiation. To test this hypothesis, we analyzed the agrp2 locus in five species from ancestral lineages that are known (Verheyen et al. 2003) to have diverged before the onset of the adaptive radiation in Lake Victoria (i.e., from Lake Kivu). The more distantly related lineages include nonstriped species from Lake Kivu (“Haplochromis gracilior) and Lake Edward (Thoracochromis pharyngalis). Two more closely related lineages include striped and nonstriped species endemic to Lake Kivu (H. vittatus and H. paucidenis) and a striped riverine haplochromine species (Astatotilapia status persii) from Kalambo River and Rusizi River (Greenwood 1979; Seehausen et al. 2003; Meier, Marques, et al. 2017), which form a connection between Lake Kivu and Lake Tanganyika. Horizontal stripes are only present in the more closely related lineages (fig. 1; Luc et al. 2001; McGee et al. 2016; Meier, Marques, et al. 2017). To test whether the causal alleles underlying stripe pattern divergence were already present in these more ancient haplochromine cichlid lineages, we analyzed the agrp2 locus in all striped ancestral species as well as in H. gracilior which was previously proposed as the source population of the Lake Victoria radiation (Verheyen et al. 2003).

First, we calculated $F_{ST}$ between striped and nonstriped phenotypes of Lake Victoria cichlids, and the ancestral lineages. From the most differentiated region (region LVRS, 538 bp, $F_{ST} = 0.88$) we built a haplotype network. To reveal the evolutionary origin of the causal variants in the Lake Victoria superflock, we added species from Lake Malawi to the haplotype network (fig. 4A). The Lake Victoria superflock is a group of 700 haplochromine cichlid species endemic to the region around Lake Victoria and nearby western rift lakes in East Africa (Meyer et al. 1990; Verheyen et al. 2003; Seehausen 2006; Genner et al. 2007). The haplotype network shows that striped and nonstriped Lake Malawi cichlids have different haplotypes than striped Lake Victoria species (fig. 4A), as already suggested by the results above (fig. 2D and E). Yet, striped Lake Victoria species share the same haplo-Christophe Batailletypo with the two striped species of the ancestral lineages of the Lake Victoria radiation (riverine A. stappersii and H. vittatus from the older Lake Kivu). We can, therefore, conclude that the cis-regulatory interval (fig. 2A and B) must have evolved after their split from their common ancestor with Lake Malawi cichlids (2–4 Ma) but before their major radi- tion into the endemic species flocks of Lake Victoria and Lake Kivu (>0.5 Ma, the age of Lake Kivu; Verheyen et al. 2003). To identify the ancestral lineage from which the haplotype at the agrp2 locus originated, we used ChromoPainter (Lawson et al. 2012). For several Lake Victoria species, we calculated the per-site probability of ancestry along haplotypes (fig. 4B). Species from Lake Victoria acted as recipients with three ancestral striped and nonstriped species acting as donors (i.e., ancestral haplotypes that are sources of recipient haplotypes). In total, we used three striped and nonstriped recipient species each and ran two separate analyses for every striped and nonstriped recipient haplotype. Apart from the three ancestral lineages that acted as donors in the first analysis, all striped species had one nonstriped within-lake donor acting as a control, whereas all nonstriped species had one striped within-lake donor as control. Thereby, four donor species were competing in every analysis where a single recipient haplotype was tested—three ancestral donors and one within-lake control. If the cis-regulatory interval would have evolved within Lake Victoria, we would expect high per-site
The evolutionary origin of the major effect haplotype of Lake Victoria stripe divergence predates the adaptive radiation of Lake Victoria. (A) Haplotype network of the highest $F_{ST}$ region (LVRS, 538 bp) between striped and nonstriped species from the Lake Victoria superflock shows the stripe and nonstripe haplotype. Mismatches represent heterozygous individuals, that is, not all species that were assigned a stripe phenotype are homozygous for the stripe haplotype. Interestingly, these species do not show a complete stripe pattern consisting of a dorsolateral and a midlateral stripe but either show only one stripe (Haplochromis paropius) or represent a striped individual of a species (H. phythophagus) that displays a polymorphism by presence/absence of horizontal stripes. After $F_{ST}$ calculation, we included the Lake Malawi mbuna which do not share the Lake Victoria stripe haplotype. However, the riverine Astatotilapia stappersii, the striped species from Lake Kivu (H. vittatus), and all striped Lake Victoria cichlids share the stripe haplotype. (B) ChromoPainter analysis visualizes the probability of haplotypic segments to be shared among
probability of ancestry for these within-lake comparisons—this is not the case. This result underscores the role of old standing genetic variation from ancestral lineages in driving the repeated evolution of stripes in the Lake Victoria cichlid species flock. We found strong evidence that the cis-regulatory interval in striped Lake Victoria species (candidate regions LV 1 and LV 2; fig. 2B) is most closely related to the riverine A. stappersii whereas other segments of the agrp2 locus are more closely related to the striped species from Lake Kivu (H. vittatus; fig. 4B).

The region of highest differentiation between striped and nonstriped species of the whole Lake Victoria superflock (fig. 4A) overlaps with candidate region LV 1. This region shows a higher probability of ancestry from the striped donor species of the riverine haplochromine (A. stappersii) than from all other striped donors. Pairwise comparisons between the striped and nonstriped species from Lake Victoria as well as its sister lineages revealed the same highly differentiated SNPs (fig. 3). Therefore, incomplete lineage sorting due to ancestral standing genetic variation that was introduced into the lake by the haplochromine founders is the most parsimonious explanation for the recurrent evolution of stripes in the Lake Victoria species flock.

Several recent studies across a wide range of study systems suggested that rapid speciation often involves “old genetic variants” upon which selection can act (Elmer and Meyer 2011; Machado-Schiaffino et al. 2017; Han et al. 2017; Meier, Marques, et al. 2017; Van Belleghem et al. 2017; Cameron and Whitfield 2019; Edelman et al. 2019; Jiggins 2019; Lewis et al. 2019; Marques et al. 2019; Kautte et al. forthcoming). By a comprehensive analysis of the “stripe locus” with its well-resolved genotype–phenotype connection, we provide additional insights into how ancestral standing genetic variation at the root of adaptive radiations can facilitate rapid phenotypic divergence within species flocks.

By tracing the evolutionary history of highly associated variants, our study sheds light on the origin of the genetic basis of horizontal stripes, an adaptive phenotype that evolved repeatedly within the hundreds of species of the East African cichlid radiations (Seehausen and Alphen 2001). Our findings show how different cis-regulatory regions of the same gene, agrp2, underlie rapid phenotypic divergence in the adaptive radiations of haplochromine cichlid fishes. We discovered that ancestral variants that form the genetic basis for stripe phenotypes in the Lake Victoria radiation predate the lake colonization and were introduced into it by the ancestors of this species flock and thereby allowed the repeated gain and loss of horizontal stripes within <100,000 years. In this radiation of more than 500 species, ancestral variants with an identified phenotypic effect (Kratochwil et al. 2018) permitted the repeated phenotypic diversification and explosive speciation that characterizes the Lake Victoria cichlid fish adaptive radiation.

Materials and Methods

Experimental Model and Sampling Details
This study was performed in accordance with the rules of the animal research facility (T-16/13) of the University of Konstanz and the animal protection authorities of the State of Baden-Württemberg.

We obtained whole-genome resequencing data from mostly wild-caught individuals from several different sources (see supplementary table S1, Supplementary Material online). The whole-genome samples from Malinsky et al. (2018), Meier, Sousa, et al. (2017), Meier, Marques, et al. (2017), and McGee et al. (2016) were obtained from wild-caught individuals. The genomes from Brawand et al. (2014) were inbred individuals from different laboratories.

Additionally, we sequenced 38 samples using target enrichment and 15 samples using whole-genome resequencing. These samples were obtained from wild-caught individuals from commercial breeders and maintained in the animal research facility of the University of Konstanz.

For most species, we sampled one individual per species to obtain a comprehensive data set that includes all major lineages. The reason why we have two samples of some species is that we used a combination of available genomes and target enrichment data for the agrp2 locus. This also allowed us to verify that different sequencing approaches did not introduce any biases during the downstream analyses (see supplementary table S1, Supplementary Material online).

Generally, we were very conservative with the phenotyping and classified a species as “striped” if either the male or the female possessed a horizontal stripe along the lateral side. All specimens that we sampled for this study were phenotyped using a photography chamber as described in Kratochwil et al. (2018). To the best of our knowledge, there are few polymorphic species and these exceptions (Neolamprologus buescheri, H. phythophagus) we sampled ourselves and documented the phenotype accordingly. In our analyses, these samples appear separately as if belonging to a striped and a nonstriped species.

Species Names and Assignment
Several of the analyzed species have different names across the literature. Because no commonly accepted taxonomy of
cichlids is available, we added a column “Current Status” in supplementary table S1, Supplementary Material online, which is based on the current classification of the “Catalog of Fishes of the California Academy” (Fricke 2020).

Target Enrichment Data
Target enrichment data were produced using customized 120-nt baits with ~3× flexible tiling density by MYBaits for the 270-kb interval around agrp2. Baits were designed from the P. nyererei reference genome (Brawand et al. 2014) which was curated by filling gaps of the genome assembly using Sanger sequencing reads so that the modified genome now contains the “stripe interval” (Kratochwil et al. 2019). This version of the P. nyererei genome is available in Dryad: https://doi.org/10.5061/dryad.bnz57h467 (Kratochwil et al. 2020).

DNA was extracted either from muscle tissue or from fin clips stored in EtOH following the DNeasy Blood & Tissue Protocol (QIAGEN) or Genaxxon Genomic DNA Purification Mini Spin Column Kit (Genaxxon Bioscience GmbH), respectively. For library preparation, we used the Illumina TruSeq Nano HT Library Preparation Kit (Illumina Inc.) following the manufacturer’s guidelines. For the baits, we followed the MYBaits manual v3 (https://arborebiocs.com/wp-content/uploads/2017/10/MYBaits-manual-v3.pdf, last accessed September 25, 2020) and hybridized the probes at 65°C for 22 h. Probes were sequenced in paired-end mode on a HiSeq 2500 system.

Whole-Genome Resequencing
DNA was extracted as explained above and DNA concentration was measured with fluorescence spectrophotometry by Qubit (Invitrogen). For library preparation, we used the Illumina TruSeq Nano HT Library Preparation Kit (Illumina Inc.) following the manufacturer’s guidelines. Samples were run on a Bioanalyser 12000 Chip to assure a high quality of DNA libraries and afterwards amplified using PCR. Finally, size distributions of all libraries were checked on a Bioanalyser HS chip before pooling them equimolarly. Sequencing was performed in paired-end mode (151PE) on a HiSeq X Ten platform (Illumina Inc.). Quality of the sequenced reads was assessed using MultiQC (Ewels et al. 2016).

The short-read data have been archived in the NCBI SRA database under the bioproject accession number PRJNA649899.

Quality Control and Statistical Analysis
Illumina adapters were trimmed from the raw fastq reads using picard v2.17.11 (http://broadinstitute.github.io/picard, last accessed September 25, 2020), and reads were mapped to the curated P. nyererei reference genome (Kratochwil et al. 2019, 2020) using bwa mem v0.7.12 (Li and Durbin 2009) and duplicate reads were marked with the standard filter (“–min-mapping-quality 30 –min-base-quality 20 –min-supporting-allele-qsum 0 –genotype-variant-threshold 0”) and population options (population-based Bayesian inference model) in freebayes v1.1.0 (Garrison and Marth 2012). We decomposed multiple nucleotide polymorphisms in the VCF file into one SNP per line using a custom python script. The resulting VCF file was then hard-filtered using common hard filters from vcfLib’s vcfFilter (“QUAL > 1 & QUAL/AO > 10 & SAF > 0 & SAR > 0 & RPR > 1 & RPL > 1”), where “QUAL” refers to the quality of the variant site and thus removes really bad sites; “QUAL/AO > 10” requires an additional contribution of each alternative allele observation of 10 log units (~Q10 per read); “SAF > 0 & SAR > 0” requires that alternative allele observations are present on both strands; and “RPR > 1 & RPL > 1” requires that at least two reads with alternative allele observations are placed toward each side of the variant site. Additionally, we used VCFTools v0.1.15 (Danecek et al. 2011) to remove indels, include only biallelic sites (~max-alleles 2) and exclude sites that are missing in more than 5% of the samples (~max-missing 0.05).

Finally, the filtered VCF file was normalized using vt normalize (Tan et al. 2015). Mean effective sequencing depth, estimated from filtered VCF files using samtools flagstat (Li et al. 2009) can be found in supplementary table S1, Supplementary Material online.

Next, we extracted phase-informative reads using quality filters –base-quality 13 and –read-quality 10 before phasing with SHAPEIT v2.r790 (Delaneau et al. 2012) and generated individual consensus fasta sequences with a custom python script. The consensus base was only kept when the site depth was above 5× coverage.

To calculate mean absolute genetic divergence ($d_{XY}$) and mean relative genetic differentiation ($F_{ST}$) between striped and nonstriped species, we used the program Saguaro (Zamani et al. 2013). With no prior assumption about the relatedness of the species, Saguaro creates local distance matrices for each region of the genome. This method infers local relationships among individuals in the form of genetic distance matrices and assigns segments across the genomes to these topologies. Thereby, it is possible that a single SNP that is alternatively fixed/highly associated between two populations results in a candidate region for high relative genetic differentiation.

To identify regions within the stripe interval that differ between stripe phenotypes, we ran TWISSST (Van Belleghem et al. 2017) for each lake separately. To reduce computation time, we used a subset of six species per lake (three striped and three nonstriped) resulting in 15 unrooted topologies. For this, we followed the authors’ recommendations (https://github.com/simonhmartin/twisst/, last accessed September 25, 2020), which involved variant calling with GATK’s Haplotype Caller (Poplin et al. 2018), filtering with VCFTools v0.1.15 (Danecek et al. 2011), and phasing with beagle 4 (Browning and Browning 2007). Finally, we constructed neighbor-joining trees for SNP windows (window size 50) in PhyML v3.1 (Guindon et al. 2010).

The curated P. nyererei reference genome assembly (Kratochwil et al. 2019) had a 26,936-bp zero coverage assembly gap on scaffold 3 that we removed from all plots as we also did not find this region in closely related species (Maylandia zebra, A. calliptera).
Next, we inferred a gene tree for the loci with the highest $F_{ST}$ values (candidate regions LM, LV 1, and LV 2) using jModelTest v2.1.1 (Darriba et al. 2012) to find the appropriate substitution model and BEAST 2 (Bouckaert et al. 2014). To compare the topology of the gene trees with a species tree, we inferred the phylogenetic relationships using 33 genomes (fig. 1). In brief, we mapped the 33 whole-genome sequences to the Oreochromis niloticus genome assembly (NCBI: GCA_000188235.1) which is more complete (higher scaffold N50) than the P. nyererei genome. Variant calling and filtering steps were performed as described earlier. In the python script used to generate individual consensus fasta sequences, we applied a maximum missingness filter of 0.75 that excludes sites on the basis of the proportion of missing data. Then, from each genome we extracted loci with a maximum physical extent of 3,000 bp each of which a minimum of 2,000 sites had to be covered. These genome-wide loci were selected randomly, requiring a minimum distance of 100 kb between loci resulting in a total of 6,545 genome-wide loci. Next, we inferred single-gene trees of all loci using IQ-tree 1.6.9 (Nguyen et al. 2015) with the ModelFinder option (Kalyaanamoorthy et al. 2017) for automatic selection of the appropriate model of evolution and with 100 rounds of ultrafast bootstrapping (Hoang et al. 2018) and estimation of the Shimodaira–Hasegawa-like approximate likelihood ratio test (Guindon et al. 2010), respectively. Ultimately, we built the species tree using all 6,545 gene trees in ASTRAL-III (Zhang et al. 2018). All trees were illustrated with FigTree 1.4.0. To provide additional supporting evidence that the substitutions close to and within the 5’-UTR of agrp2 in Lake Malawi mbuna could have an effect on agrp2 transcription, we screened the two divergent haplotypes for TFBSs using JASPAR (Fornes et al. 2020). For this, we extracted the flanking sequences of each SNP ($\pm$10 bp) within candidate region LM (fig. 3 and supplementary fig. S2, Supplementary Material online) and screened for TFBSs above a conservative threshold of 0.85 (Kwon et al. 2012) of the relative matrix score in at least one of the two species. The difference between relative scores (delta $P_{dem}$ — $P_{cyaneorhabdos}$) of Ps. demasoni and Ps. cyaneorhabdos serves as an indicator of differential regulation of agrp2 between the nonstriped and striped species. Supplementary table S2, Supplementary Material online, gives a summary of all results as well as location of gene expression which we collected from ZFIN (https://zfin.org/, last accessed September 25, 2020).

We performed the same analysis for region LV 1 using the nonstriped species P. nyererei, Pnye, and striped H. sauvagei, Hsau. We focused on LV 1, as it had the strongest association, which donors are essential to explain the recipient. There, we used different striped and nonstriped donors (ancestral species which are the sources of admixture) and three striped and three nonstriped recipient species, which represent the recipients of admixture. For visualization, we used the pheatmap function (Kolde 2019) in R. Scripts are available in the GitHub repository (https://github.com/sabineurban, last accessed September 25, 2020).

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

Acknowledgments
This work was funded by fellowships of the International Max Planck Research School for Organismal Biology (to S.U.), the Swiss National Science Foundation (P300PA_177852 to A.N.), the Elite-Program-for-Postdocs, Baden-Württemberg Foundation, the Deutsche Forschungsgemeinschaft (DFG, KR 4670/2-1 and KR 4670/4-1 to C.F.K.), and the European Research Council (ERC Advanced Grant, GenAdap 293700 to A.M.). Computations were carried out on resources provided by the Scientific Compute Cluster of the University of Konstanz. The authors thank Jannik Beninde, Jan Gerwin, Yipeng Liang, Stefan Gerlach, and the staff of the animal research facility of the University of Konstanz for their valuable help.

Author Contributions
S.U. wrote the article with contributions from all authors; S.U. and C.F.K. conducted sample collection; S.U. did bench work; S.U. and A.N. conducted analyses; C.F.K. designed the study. C.F.K. and A.M. supervised the study.

Data Availability
Fastq raw reads of all samples sequenced for this study (supplementary table S1, Supplementary Material online) are deposited at the NCBI Sequence Read Archive under the bioproject accession number PRJNA649899.

References


Different Sources of Allelic Variation. doi:10.1093/molbev/msaa237


