Possible involvement of ghost introgressions in the striking diversity of V1R receptor genes in East African cichlids

Shunsuke Taki[1], Zicong Zhang[2], Mitsuto Aibara[1], Tatsuki Nagasawa[1], and Masato Nikaido[1*]

[1]School of Life Science and Technology, Institute of Science Tokyo, Tokyo, Japan

[2]Institute for the Advanced Study of Human Biology, Kyoto University, Kyoto, Japan

[*]Correspondence: Masato Nikaido, mnikaido@bio.titech.ac.jp

Abstract

Cichlids that have experienced adaptive radiation are genetically close but exhibit extreme ecological and morphological diversity, making them widely used as textbook examples to understand speciation mechanisms. V1R-type olfactory receptors are highly conserved at the amino acid sequence level among teleost fish and believed to play a fundamental role in reproduction. Our previous studies have reported the unexpected sequence diversity of V1Rs among some cichlid species, implying the possible involvement of V1Rs in their speciation. In this study, we extensively investigated the process of the evolutionary diversification of all six V1Rs (V1R1 to V1R6) by utilizing the genome data from 528 cichlid species, covering almost all lineages. In V1R2, two highly divergent alleles (1.17%: variant sites/CDS length) without recombination were shared and maintained among the cichlids of all East African Great Lakes. In V1R6, large number of highly variable alleles that could be derived from the multiple recombination of two highly divergent alleles (1.39%: variant sites/CDS length) were shared among the cichlids of Lake Victoria. Additionally, we also identified highly divergent alleles in V1R1 within the tribe Tropheini, and in both V1R3 and V1R6 within Trematocarini and Ectodini. Given that one of the two divergent alleles of these V1Rs emerged so rapidly during cichlid evolution, they are likely to be derived from introgression. However, the source lineages for those introgressions were not identified despite the extensive exploration implying that they have already been extinct. This work provides compelling support for the contribution of introgression to explain striking diversity of V1Rs in East African cichlids.

Keywords: adaptation, cichlids**,** ghost introgression, standing genetic variation, V1R

Introduction

Cichlids, belonging to the family Cichlidae, are a group of tropical freshwater fish that inhabit a wide range of regions from Central and South America to Africa, Madagascar, and India. The total number of species is estimated to be about 3,000, representing about 10% of all teleost fish (Kocher, 2004). Notably, about half of the extant cichlid species inhabit the East African Great Lakes (Lake Tanganyika, Lake Malawi, and Lake Victoria) (Salzburger et al., 2014), and these species have emerged through large-scale adaptive radiation in each lake (Kocher, 2004; Salzburger, 2018). These cichlids of each Lake are widely used as model organisms to investigate the mechanism of speciation because of their remarkable ecological and morphological diversity despite their genetic closeness. Cichlids are often reproductively separated by prezygotic isolation (Rometsch et al., 2020). Sexual selection and assortative mating contribute to speciation, particularly in environments where geographic isolation is less likely, such as lakes (Seehausen & Wagner, 2014).

Cichlids depend on multiple sensory systems for their reproductive behavior (Escobar-Camacho & Carleton, 2015). Research on the visual system is especially active, as cichlids exhibit a wide variety of body colors. Previous studies have shown that the visual system has diversified according to ecology and habitat, and that its environmental adaptation have contributed to speciation (Carleton & Yourick, 2020; Seehausen et al., 2008; Terai et al., 2006). The promotion of speciation through environmental adaptation of sensory systems is referred to as "sensory drive", which has been shown to play a significant role in the speciation of cichlids. The prominent example is the visual adaptation to the light environment (bule or red) through peak shift of long-wavelength-sensitive (lws) opsin together with the divergence of female preference for male nuptial coloration (bule or red) led to the speciation of Lake Victoria cichlids (Seehausen et al., 2008).

62 Cichlids use not only visual but also olfactory cues for species and individual
63 recognition (Keller-Costa et al., 2015; Santos et al., 2023). Behavioral experiments using
64 Lake Malawi cichlids have demonstrated that olfaction plays a role in conspecific recognition
65 (Plenderleith et al., 2005). The imprinting on olfactory cues through mouthbrooding have also
66 shown in Lake Victoria cichlids (Verzijden & ten Cate, 2007). In addition, compounds present
67 in the urine of Mozambique tilapia (*Oreochromis mossambicus*) act as pheromones,
68 influencing male behavior and the female reproductive system (Ashouri et al., 2023; Keller-
69 Costa et al., 2014). It is obvious that cichlids largely depend on olfaction, however, in contrast
70 to studies of the visual system, it remains unclear which olfactory receptor genes and alleles
71 contribute to species recognition and hence assortative mating.
72 Four types of olfactory receptor genes (OR, TAAR, V1R, V2R) have been identified
73 in teleost fish, all of them are expressed in the olfactory neurons (Calvo-Ochoa et al., 2021).
74 One of these, the V1R (vomeronasal type 1 receptor), which is also called ORA (olfactory
75 receptor class A related), is considered to detect pheromones in fish, as it is homologous to
76 the mammalian V1R, a known pheromone receptor (Del Punta et al., 2002; Pfister &
77 Rodriguez, 2005). Although the V1Rs are diverse in copy number and sequence in mammals,
78 they are highly conserved in teleost fish by sharing six copies of the V1R genes (V1R1 to
79 V1R6), of which the orthologs are also highly conserved at amino acid sequences level
80 (Saraiva & Korsching, 2007). Therefore, they likely to detect hormonal pheromones shared
81 across species and to contribute to the synchronization of gamete maturation and/or
82 spawning interactions (Ota et al., 2012; Policarpo et al., 2022; Saraiva & Korsching, 2007;
83 Zapilko & Korsching, 2016). The V1Rs in zebrafish has been shown to detect 4HPAA (4-
84 hydroxyphenylacetic acid) and bile acids (Behrens et al., 2014; Cong et al., 2019).
85 Additionally, V1R2 and V1R5 expressing neurons respond to male urine in cichlids

86 (Kawamura & Nikaido, 2022). Thus, V1Rs have great potential to contribute to speciation in
87 cichlids, however, it is not clear whether their genetic diversity is linked to speciation or
88 environmental adaptation, as has been shown in LWS.

89 The V1R genes of East African cichlids exhibit sequence-level diversity (Nikaido et
90 al., 2014). In our previous study, we found diverse alleles of V1R1, V1R3, and V1R6 in some
91 tribes of Lake Tanganyika cichlids. We also found two different alleles of V1R2, and revealed
92 that these are shared across species from different lakes. These alleles are separated by
93 more than 10 nonsynonymous substitutions, indicating the signature of functional
94 differentiation. Hence, V1Rs are ideal research targets for examining the possibility of
95 olfaction-mediated speciation of cichlids (Santos et al., 2023). However, the small number of
96 species analyzed hindered a complete understanding of the origin and formation process of
97 these divergent alleles.

98 In recent years, the release of whole-genome sequencing (WGS) data for more than
99 500 cichlid species has enabled comprehensive comparative genomic analyses across
100 species (Malinsky et al., 2018; McGee et al., 2020; Ronco et al., 2021). This allowed us to
101 extensively explore the V1R sequences to understand how the divergent alleles were
102 originated and maintained across lineages in morphologically and ecologically diverse
103 cichlids. In the present study, we conducted a comprehensive analysis of V1R allelic diversity
104 by using WGS data from 528 species and 908 samples of cichlids. Furthermore, we
105 investigated the potential impact of sequence-level differences on V1R function, particularly
106 ligand reception.
107

Materials and Methods

V1R gene sequences

We used short-read data from 28 tribes, 528 species, and 908 samples of cichlids registered in the NCBI Sequence Read Archive (SRA) (Figure S1, Table S1). Low-quality reads were removed using fastp v.0.23.2 (Chen et al., 2018). The filtered reads were mapped to the reference genome (Supplementary Information) using bwa-mem2 v.2.2.1 (Vasimuddin et al., 2019), and sorted and indexed with samtools v.1.15 (Danecek et al., 2021). The V1R gene regions were extracted from the sorted BAM files using samtools v.1.15 (Danecek et al., 2021). Variants were called from the extracted BAM files using GATK v.4.3.0 (McKenna et al., 2010) with the HaplotypeCaller tool. The first round of hard filtering for SNPs and INDELs was applied using GATK v.4.3.0 (McKenna et al., 2010) with the VariantFiltration tool and the following parameters: SNPs (QD<2.0, QUAL<30.0, SOR>4.0, FS>60.0, MQ<40.0, MQRankSum<-12.5, ReadPosRankSum<-8.0) and INDELs (QD<2.0, QUAL<30.0, FS>200.0, SOR>10.0, ReadPosRankSum<-20.0). Based on the resulting VCF file, quality recalibration of the BAM files was performed using GATK v.4.3.0 (McKenna et al., 2010) with BaseRecalibrator and ApplyBQSR. The second round of hard filtering was conducted as in the first round. The filtered VCF files were used to create V1R gene sequences in fasta format using bcftools v.1.15 (Danecek et al., 2021). Sequences containing heterozygous variants were phased using HapCUT2 v.1.3.3 (Edge et al., 2017). We built a custom pipeline from downloading short-read data to getting V1R gene sequences (https://github.com/taki-sh/RegionCall).

Phylogenetic analysis

To reduce the computational load of the phylogenetic analysis, we removed duplicated

132 sequences and then randomly extracted a total of 45 sequences from 337 V1R2 sequences,
133 creating five datasets (Figure 1, Figure S2-6). We performed multiple sequence alignment
134 for all of these datasets using MAFFT v.7.520 (Katoh & Standley, 2013) with the options: --
135 maxiterate 16 --localpair. Maximum likelihood trees were constructed using RAxML-NG
136 v.1.2.0 (Kozlov et al., 2019) with 1000 bootstrap replicates. The model used for phylogenetic
137 tree construction, HKY+G4, which was selected by Modeltest-NG v.0.1.7 (Darriba et al.,
138 2020). We used Dataset 1 in Figure 1, as Datasets 1 through 5 demonstrated similar
139 topological patterns (Figure S2-6). For V1R1 of Tropheini, we used 43 non-duplicated V1R1
140 sequences along with the outgroup *Oreochromis niloticus* V1R1 sequence, totaling 44
141 sequences (1000 bootstrap replicates, model: HKY+I). For V1R3 of Trematocarini, we used
142 11 non-duplicated V1R3 sequences along with the outgroup *Oreochromis niloticus* V1R1
143 sequence, for a total 12 sequences (1000 bootstrap replicates, model: HKY+I). For V1R3
144 and V1R6, we used 53 non-duplicated V1R3 sequences (100 bootstrap replicates, model
145 V1R3: HKY+G4, V1R6: GTR+G4). For visualization, we used the R packages: treeio (Wang
146 et al., 2020), ggtree (Yu et al., 2017), readr (Wickham H, Hester J, Bryan J, 2024), dplyr
147 (Wickham H, François R, Henry L, Müller K, Vaughan D, 2023), ggplot2 (Villanueva & Chen,
148 2019), ggmsa (Zhou et al., 2022), and iTOL v6 (Letunic & Bork, 2024). To calculate the
149 frequency of each allele of V1R1 and V1R2 across cichlid lineages, we used unphased
150 sequences (V1R1: 908 sequences, V1R2: 907 sequences). We used 1309 sequences for
151 the classification of V1R6 haplotypes. Additionally, to reveal the sequence diversity of V1R
152 genes in ray-finned fish, we used V1R sequences from 102 species of ray-finned fish
153 (Policarpo et al., 2022) and phylogenetic tree data (Chang et al., 2019; Rabosky et al., 2018).
154 We generated a pairwise similarity matrix for the 102 species using CIAlign v.1.0.17
155 (Tumescheit et al., 2022) and calculated the average similarity. We also calculated pairwise

156 dN/dS ratios using the CODEML program in PAML v.4.10.6 (Yang, 2007) and computed the
157 average value.
158
159 **Selection analysis**
160 Amino acid sequences translated using Aliview (Larsson, 2014), were aligned in MAFFT
161 v.7.520 (Katoh & Standley, 2013) with the options: --maxiterate 16 --localpair. This multiple
162 alignment of amino acid sequences was converted into a codon alignment using PAL2NAL
163 v.14 (Suyama et al., 2006). For branches where an accumulation of nonsynonymous
164 substitutions was observed, we estimated dN/dS ratios using branch model with the codeml
165 program in PAML v.4.10.7 (Yang, 2007) and performed a likelihood ratio test between the
166 foreground and background branches to examine the statistical significance for the elevation
167 of dN/dS in foreground branch. Clade I and Clade II were defined within V1R2. Using PAML,
168 the dN/dS ratio was calculated across all five datasets (Dataset 1 to 5, Figure S2-6), and
169 ancestral sequences were inferred for the ancestral and descendant nodes of each branch.
170 A comparison of the inferred ancestral sequences identified 11 nonsynonymous substitutions
171 common to all datasets, which were classified as mutations distinguishing Clade I and Clade
172 II. In V1R1, we also defined Clade I and Clade II based on the phylogenetic tree of the V1R1
173 gene in Tropheini (Figure S9). Similarly, we identified 11 nonsynonymous substitutions, which
174 were classified as mutations distinguishing Clade I and Clade II.
175
176 **Protein modeling and ligand-binding predictions**
177 We modeled 3D structures of V1R1, V1R2, and V1R6 using AlphaFold v.2.3 (Jumper et al.,
178 2021). Based on the most reliable structures, we estimated the coordinates of the ligand-
179 binding pockets using Deepsite (Jiménez et al., 2017) in PlayMolecule Viewer (Torrens-

180 Fontanals et al., 2024). Furthermore, we performed ligand docking simulations using Webina
181 v.1.0.5 (Kochnev et al., 2020) with the obtained coordinates. For these simulations, we used
182 4HPAA, a known ligand for zebrafish V1R2 (Behrens et al., 2014), and LCA (lithocholic acid),
183 a known ligand for zebrafish V1R1 and V1R6 (Cong et al., 2019). As a result, five relative
184 positions (poses) of the ligand in V1R1 (clade I: 3 poses, clade II: 2 poses), 18 poses in V1R2
185 (clade I: 9 poses, clade II: 9 poses), and 18 poses in V1R6 (ancestral allele: 9 poses, derived
186 allele: 9 poses) were predicted. Residues positioned within 5Å of all these poses were
187 designated as ligand-related sites. In addition, we predicted transmembrane regions using
188 DeepTMHMM v.1.0.39 (Hallgren et al., 2022). Visualization of the results was carried out
189 using ChimeraX v.1.6 (Goddard et al., 2018).
190
191 **Recombination rate**
192 We used the recombination rate of *Pundamilia nyererei* (Feulner et al., 2018) and the
193 genomic position of the V1R genes to relatively compare the recombination rates in the
194 regions surrounding the V1R genes. We obtained the genomic positions of the V1R genes
195 using BLASTN with the *Pundamilia nyererei* genome assembly (P_nyererei reference v2)
196 (Feulner et al., 2018) as the reference and *Oreochromis niloticus* V1R sequences (V1R1:
197 AB671405.1, V1R2: AB670433.1, V1R3: AB704581.1, V1R4: AB704618.1, V1R5:
198 AB704661.1, V1R6: AB704705.1 from NCBI) as the query. The genomic positions of the V1R
199 genes were identified as follows: V1R1 (chr13:8171366-8172331), V1R2 (chr13:8175817-
200 8176758), V1R3 (chr13:18527399-18530109), V1R4 (chr13:18531988-18532072), V1R5
201 (chr2:41585178-41586302), V1R6 (chr7:27107196-27108122).
202
203

204 **Results**

205 **Two highly divergent alleles of V1R2 in cichlids of East African Great Lakes**

206 We constructed a phylogenetic tree of V1R2 gene using sequences from 44 species selected
207 from all cichlid lineages (Figure 1). Because cichlids in the East African Great Lakes have
208 independently undergone adaptive radiation in each lake, species inhabiting the same lake
209 are closely related, with Lake Malawi and Lake Victoria each forming monophyletic groups
210 (Svardal et al., 2021). However, the topology of the V1R2 gene tree did not match the species
211 tree and was divided into clade I and clade II with maximal bootstrap probability (Figure 1).
212 Clade I included cichlids from the outgroup (India, Madagascar), African Riverine, Lake
213 Tanganyika, Lake Malawi, and Lake Victoria. Clade II included cichlids from Lake Tanganyika,
214 ancient Haplochromini, modern Haplochromini, Lake Malawi, Upper Nile lineage, Eastern
215 lineage, Congolese lineage, and Lake Victoria. Notably, both clade I and clade II contained
216 lineages from all of the three East African Great Lakes. Clade I and clade II were separated
217 by 11 amino acids (Figure 1, see Method). In addition, the dN/dS ratio of the branch was
218 significantly higher than that of the background (Table S2). This elevation of dN/dS indicates
219 that positive selection operated on the branch where clade II diverged from clade I.

220 We calculated the frequencies of the two major alleles (clade I and clade II) of V1R2 in
221 each cichlid lineage (Figure 2). In cichlids from Lamprologini and Ectodini, inhabiting Lake
222 Tanganyika, V1R2 alleles were fixed into clade II. In Haplochromini from Lake Malawi and
223 Lake Victoria, the alleles were not fixed into neither allele, becoming the mixtures of
224 homozygotes of clade I or clade II, as well as their heterozygotes. Despite the extensive
225 exploration, clade II was not observed in the lineages from India, Madagascar, New World,
226 or African Riverine. Interestingly, no recombination between the two alleles was detected in
227 any lineage in spite that the alleles co-exist as polymorphisms within population or species

228 with only one exception for Cyphotilapiini. The polymorphic sites separating clade I and clade
229 II in cichlids are highly conserved among non-cichlids ray-finned fish, of which V1R are all
230 classified into clade I (Figure S7).
231
232 **Highly variable alleles of V1R6 with multiple recombination in cichlids of Lake Victoria**
233 We next explored the V1R6 gene and constructed the phylogenetic tree, of which the
234 topology is consistent with the species tree (Figure 3A). Interestingly, we identified striking
235 number of different alleles that were separated by the combination of 13 variant sites in Lake
236 Victoria cichlids (Figure 3B). These variable alleles were shared among species as
237 polymorphisms. It is highly unlikely that the substitutions occur at the same site multiple times
238 independently in each allele. Therefore, we hypothesized that the striking number of different
239 alleles were derived from the recombination between two highly divergent alleles present in
240 the ancestral population in Lake Victoria (Figure 3C). Indeed, we calculated the
241 recombination rate in the genomic region surrounding V1R6 using the data from previous
242 study (Feulner et al., 2018), showing the higher recombination rate compared to other V1R
243 surrounding regions (Figure S8). Based on this hypothesis, we focused on the 13 variant
244 sites identified in Lake Victoria and classified the V1R6 alleles of the cichlids of all East
245 African Great Lakes and rivers. In cichlids inhabiting outside the Lake Victoria basin, we
246 found no apparent signature of recombination between highly divergent alleles. We then
247 reconstructed the ancestral sequences at the node marked with an arrowhead in Figure 3A
248 and defined it as “ancestral allele” (Figure 3B left top, dark blue nucleotide). Next, by
249 subtracting the nucleotides of “ancestral allele” (dark blue) from the 13 polymorphic sites in
250 Lake Victoria, we reconstructed a hypothetical “derived allele” (yellow), which could be a
251 counterpart sequence of the past recombination (Figure 3B right top). As a result, the alleles

252 found in Lake Victoria are divided into the “ancestral”, “derived” and their “mixtures”. Notably,
253 the cichlid of Upper Nile, Eastern lineage, and Congolese lineage also possess recombinant
254 alleles similar to those in Lake Victoria (Figure 3B). However, in spite of the extensive
255 exploration, the hypothetical “derived allele” was not identified in extant cichlids except for
256 Lake Victoria (e.g., haplotypes Lake Victoria 51, 52 in Figure 3B right bottom).
257

258 **Highly divergent alleles of V1R1, V1R3, V1R6 in cichlids of particular tribes**

259 We identified two highly divergent alleles named clade I and clade II, which are separated by
260 11 amino acids, in the V1R1 of Tropheini (Figure 4A, see Method). The monophyly of clade
261 I and clade II are supported by maximal bootstrap probability (Figure S9). Given that the
262 amino acid sequence of clade I of Tropheini is mostly similar with other cichlids, clade I is
263 defined as “ancestral” and II as “derived” alleles. Furthermore, the dN/dS ratio of the branch
264 was significantly higher than that of the background (Table S2). This elevation of dN/dS
265 indicates that positive selection operated on the branch where clade II diverged from clade I.
266 Next, we calculated the frequencies of the two major alleles (clade I and clade II) in the V1R1
267 of Tropheini. We found that 28% of samples in Tropheini possess clade I, while 65% possess
268 clade II (Figure 4B). We extensively explored the clade II sequences, but did not find them in
269 any cichlids other than Tropheini. The polymorphic sites separating clade I and clade II in
270 cichlids are highly conserved at clade I among ray-finned fish (Figure S10).

271 We also identified highly divergent allele in the V1R3 in cichlids of Ectodini and
272 Trematocarini. Seven nonsynonymous substitutions were observed on the branch
273 representing the common ancestor of Ectodini (Figure S11). The dN/dS ratio of this branch
274 was significantly higher than that of the background (Table S2). In Trematocarini, ten
275 nonsynonymous substitutions were observed on the branch representing the common

276 ancestor of Trematocarini, excluding *Trematocara unimaculatum* (Figure S12). The dN/dS
277 ratio of this branch was significantly higher than that of the background (Table S2). In V1R6,
278 we found the divergent alleles in Ectodini and Trematocarini (Figure S13) and the dN/dS ratio
279 of those clades (Ectodini: 2.69, Trematocarini: 1.61) were significantly higher than that of the
280 background (Table S2).
281
282 **V1R2 allele sites are located near ligand binding pocket**
283 To elucidate the functional impact of the amino acid substitutions observed between the two
284 major divergent alleles of V1R2, we predicted the 3D structure and ligand binding sites of
285 V1R2 (Figure 5A). First, we predicted the 3D structures for each allele using the amino acid
286 sequences of clade I and clade II. We then used 4HPAA, which is one promising candidate
287 of the ligand (Behrens et al., 2014; Kawamura & Nikaido, 2022), to estimate the ligand
288 binding sites and compared them between the alleles. As a result, we identified that eight out
289 of the 11 amino acid substitution sites (TM2: C62→Ser, ECL1: Q65→Pro, TM4: A145→Ser,
290 A146→Thr, ECL2: P159→Ser, H169→Leu, TM5: V185→Ile, TM6: D247→Glu) are located
291 close to the predicted ligand pocket (Figure 5B). Furthermore, four of these sites (TM4: A145
292 →Ser, A146→Thr, TM5: V185→Ile, TM6: D247→Glu) coincided with the ligand binding sites
293 (Figure 5B). The skewed amino acid substitutions of V1R2 towards the ligand binding sites
294 were statistically significant (Fisher's exact test: p=0.0389).
295 We also the same analysis of V1R1 and V1R6 (Figure S14A, S15A). In V1R1, we identified
296 that five out of the 11 amino acid substitution sites (TM2: V59→Gly, ECL2: I164→Leu, TM5:
297 G188→Val, TM7: S273→Thr, G280→Ser) are located close to the predicted ligand pocket
298 (Figure S14B). Furthermore, three of these sites (ECL2: I164→Leu, TM7: S273→Thr, G280
299 →Ser) coincided with the ligand binding sites (Figure S14B). However, the skewed amino

300 acid substitutions of V1R1 towards the ligand binding sites were not statistically significant

301 (Fisher's exact test: p=0.0922). In V1R6, we could not identify any amino acid substitution

302 sites that are located close to the predicted ligand pocket (Figure S15B).

303 **Discussion**

304 **Allelic and functional diversity of V1R in the East African cichlids**

305 We conducted large-scale analyses focusing on the diversity of V1R in cichlids, using WGS
306 data from 528 species and 908 samples, which recently became available by the large scale
307 comparative genomic studies (Malinsky et al., 2018; McGee et al., 2020; Ronco et al., 2021).
308 As a result, we identified an accumulation of nonsynonymous substitutions and the
309 signatures of the positive selection in V1R1, V1R2, V1R3, and V1R6 (Figure 6A). Notably,
310 two highly divergent alleles were found in V1R1 and V1R2 of the cichlids of Tropheini (Fig.
311 4B) and those inhabiting the East African Great Lakes (Figure 2B), respectively. In addition,
312 highly divergent alleles were found in V1R3 and V1R6 of the cichlids of Trematocarini and
313 Ectodini (Figure S11-13). The results of the present study well supports previous findings
314 from small scale species sampling (Nikaido et al., 2014).

315 Each of the V1R gene orthologs is highly conserved across fish species (Ota et al., 2012;
316 Policarpo et al., 2022; Saraiva & Korsching, 2007; Zapilko & Korsching, 2016). The recently
317 published V1R gene sequences of ray-finned fish (Policarpo et al., 2022) also exhibit high
318 sequence similarity and low dN/dS ratios (Figure S16). Furthermore, the polymorphic sites
319 separating the divergent alleles of V1R1 and V1R2 in cichlids are highly conserved among
320 ray-finned fish (Figure S7, S10). These results suggest that the diversification of the V1R
321 gene is unique to cichlids among ray-finned fish and are likely to be linked with their rapid
322 adaptive radiation.

323 The prediction of the 3D structure of V1R2 and the estimation of the sites interacting
324 with the ligand revealed that the amino acid residues distinct between clade I and II are
325 located close to the ligand pocket. In addition, the amino acid substitutions were skewed
326 towards the ligand binding sites, which was statistically significant (Figure 5B). Two amino

327 acid substitutions occurred in ECL2 (extracellular loop 2) (Figure 5B), which is involved in
328 ligand selectivity in class A GPCRs, a category that includes V1R (Nicoli et al., 2022). In the
329 bitter taste receptor T2R, which is phylogenetically close to V1R, ECL2 also contributes to
330 ligand binding (Xu et al., 2022). These lines of observation suggest that two divergent clades
331 of V1R2 are likely to be distinct each other in terms of ligand binding.

332 V1R6 has the lowest sequence similarity and the highest dN/dS ratio among six V1R
333 genes across all ray-finned fish (Figure S16). Hence, the pressure of negative selection
334 operated on V1R6 is weaker than other V1Rs not only in cichlids but also across ray-finned
335 fish. It is worth noting that V1R6 in Lake Victoria basin is highly variable with number of
336 different alleles due to the multiple recombination between divergent alleles (Figure 3B, 3C).
337 The recombination rate in the V1R6 region of Lake Victoria cichlids is notably high (Figure
338 S8), indicating that V1R6 may be located in a recombination hotspot. This is in sharp contrast
339 with V1R1 and V1R2, in which recombination is strictly suppressed. Thus, the structural
340 constraint on V1R6 seems to be weak, implying that the ligand specificity of V1R6 may be
341 lower than other V1Rs. In the 3D structural analysis, a ligand pocket of V1R1 and V1R2 was
342 predicted within the TM region, that is similar to typical class A GPCRs (Figure 5B, S14B). In
343 contrast, the ligand binding of V1R6 was predicted to be mediated by the ECL2 region (Figure
344 S15B). This suggests that V1R6 may be distinct with V1R1 and V1R2 in terms of functions
345 and ligand binding mechanisms. Further verification by considering the distinctness of the
346 predicted structural differences among V1Rs will provide insight into the function of V1Rs in
347 fish, which remain a topic of ongoing debate (Behrens et al., 2014; Cong et al., 2019; Saraiva
348 & Korsching, 2007).

349

350 **The origin of striking V1R diversity through ghost introgression**

351 It is of primary importance to understand how the remarkable diversity of V1Rs has been
352 originated and retained among the cichlids over the 10 million years of their evolution. We
353 first investigated the origin and process of V1R2 diversification by extensively examining the
354 existence of clade I and clade II alleles in each lineage of cichlids (Figure 2). As a result,
355 clade II was found in cichlid lineages inhabiting the East African Great Lakes, and was not
356 found in lineages that diverged prior to the Lake Tanganyika radiation. Specifically, the
357 cichlids of Lamprologini and Ectodini were fixed into clade II, and those of other Lake
358 Tanganyika tribes were fixed into clade I. The cichlids of Haplochromini were homozygotes
359 of clade I or clade II as well as the heterozygotes of them. The pattern of V1R2 allele
360 distribution suggest that the clade II suddenly appeared at the timing of radiation of Lake
361 Tanganyika cichlids around 10 million years ago. Given that the adaptive radiation in Lake
362 Tanganyika was so rapid, it is unlikely that 12 substitutions would accumulate step by step in
363 a single allele during that time, regardless of the operation of positive selection. Therefore, it
364 is parsimonious to assume that the clade II allele has already existed in some
365 populations/species prior to the radiation, and was introgressed into Lake Tanganyika
366 through hybridization.

367 Several examples of standing genetic variation (SGV), some of which date back to even
368 before the adaptive radiations in Lake Tanganyika (Nakamura et al., 2023), have been
369 identified in the genomes of East African cichlids (Nakamura et al., 2021; Svardal et al., 2021).
370 Clade I and II alleles of V1R2 well represents an example of such old SGV. The Melting-pot
371 hypothesis is one of the most plausible scenarios to explain the origin of old SGV (Nakamura
372 et al., 2023; Weiss et al., 2015). According to the Melting-pot hypothesis, several cichlid
373 populations/species, which were scattered in wetlands and small lakes, hybridized just before
374 the adaptive radiation in Lake Tanganyika. In case that each of the scattered populations

375 were fixed into clade I and clade II, the hybridization of them may provide the origin of SGV
376 of highly divergent V1R2. However, one of the two alleles comprising the SGV of V1R2, clade
377 II, were not found in spite of an extensive exploration of 46 basal lineages of riverine species
378 (Figure 2). The absence of clade II allele in the riverine species suggest that the transporter
379 lineage of clade II alleles have been already extinct.

380 In recent years, gene flow from extinct lineages (ghost introgression) has been detected
381 in various vertebrates from gobiid fish to elephants (Ai et al., 2015; Frei et al., 2022;
382 Gopalakrishnan et al., 2018; Kato et al., 2023; Palkopoulou et al., 2018; Zhang et al., 2019).
383 In cichlids, sudden emergence of clade II of V1R2 in Lake Tanganyika cichlids can be
384 explained by ghost introgression. In general, introgressed alleles are detected as highly
385 divergent compared with non-introgressed genomic region. However, neutral genomic
386 regions are often subject to recombination and drift after introgression, making it difficult to
387 detect the signature of introgression. In this regard, V1R2 is a good example to detect
388 introgression because recombination was strictly restricted. African riverine lineages that
389 existed prior to the adaptive radiation are referred to as "transporters" of genetic diversity
390 (Loh et al., 2013). Notably, gene flow from Steatocranini, a riverine lineage that diverged
391 before the adaptive radiation (Irisarri et al., 2018), and the appearance of clade II of the V1R2
392 occurred at identical time period. Thus, it is highly likely that the remarkable diversity of V1R2
393 alleles (clade II) derived from the introgression from an extinct African riverine lineage that
394 diverged prior to the adaptive radiation (ghost introgression) or from the unexamined lineages
395 (Figure 6B).

396 Our data on V1R6 suggest that remarkable number of different alleles derived from the
397 multiple times of recombination of two highly divergent alleles (ancestral and derived)
398 separated by 13 nucleotide substitutions was observed in cichlids of Lake Victoria basin

399 (Figure 3B). In Lake Victoria, the recombination of the LWS opsin gene was observed, which
400 was explained by the hybridization between the Congo and Upper Nile lineages (Meier et al.,
401 2017). In contrast, given that both the Congo and Upper Nile lineages also exhibit
402 recombinant sequences, recombination between ancestral and derived alleles of V1R6
403 occurred earlier than the hybridization event of the Congo and Upper Nile lineages. However,
404 none of the extant species in Haplochromini, except for Lake Victoria cichlids, has the derived
405 allele. Considering that the estimated mutation rate in cichlids is to be 3.5×10^{-9} (per site per
406 generation), it is unrealistic that 13 substitutions accumulate during the short period of the
407 formation of the cichlids of Lake Victoria basin (~400,000 years ago) (Johnson et al., 1996;
408 Salzburger et al., 2014). Therefore, the derived allele of V1R6 was also likely to be
409 introgressed from an extinct lineage (Figure 6B). Similar to the logics of V1R2 and V1R6, the
410 highly divergent alleles observed in V1R1 of Tropheini, as well as the in V1R3 and V1R6 in
411 Ectodini and Trematocarini, are also likely to have been derived from ghost introgression.
412 It is worth noting that the nucleotide sequence differences between the two alleles found
413 in V1R1, V1R2, and V1R6 were mostly all approximately higher than 1% (V1R1 = 1.14%,
414 V1R2 = 1.17%, V1R6 = 1.39%). It is possible to hypothesize that each divergent allele has
415 been introduced into the ancestral population through a single ghost introgression event
416 (Melting-pot), after which the derived alleles independently sorted to each lineage.
417 Alternatively, introgression events in each lineage may have led to the introduction of the
418 derived alleles in V1R1, V1R2, and V1R6 (Figure 6A, 6B).

419 **Conclusion**

420 In this study, through a comprehensive analysis of the V1R genes covering the most lineages
421 of East African cichlid species, we identified the large scale standing genetic variation that
422 could be explained by ghost introgression. It will be necessary to conduct the genome-wide
423 statistical analyses rather than at the gene level to examine the ghost introgression and
424 Melting-pot scenarios in further research. We also revealed differences in the structural
425 constraint and selection pressure among different V1Rs in cichlids, implying the functional
426 distinctness among them. This research provides important insights into the possibility of
427 V1R mediated speciation and the genetic source enabling the adaptive evolution of cichlids.
428

429 **Acknowledgements**

430 This study was supported by JSPS KAKENHI (20KK0167, 24K02074, 24KK0141) to MN.

431 Phylogenetic analyses were partially performed using the NIG supercomputer at ROIS

432 National Institute of Genetics.

433

References

Ai, H., Fang, X., Yang, B., Huang, Z., Chen, H., Mao, L., Zhang, F., Zhang, L., Cui, L., He, W., Yang, J., Yao, X., Zhou, L., Han, L., Li, J., Sun, S., Xie, X., Lai, B., Su, Y., … Huang, L. (2015). Adaptation and possible ancient interspecies introgression in pigs identified by whole-genome sequencing. *Nature Genetics*, *47*(3), 217–225. https://doi.org/10.1038/ng.3199

Ashouri, S., Da Silva, J. P., Canário, A. V. M., & Hubbard, P. C. (2023). Bile acids as putative social signals in Mozambique tilapia (Oreochromis mossambicus). *Physiology & Behavior*, *272*, 114378. https://doi.org/10.1016/j.physbeh.2023.114378

Behrens, M., Frank, O., Rawel, H., Ahuja, G., Potting, C., Hofmann, T., Meyerhof, W., & Korsching, S. (2014). ORA1, a zebrafish olfactory receptor ancestral to all mammalian V1R genes, recognizes 4-hydroxyphenylacetic acid, a putative reproductive pheromone. *The Journal of Biological Chemistry*, *289*(28), 19778–19788. https://doi.org/10.1074/jbc.M114.573162

Calvo-Ochoa, E., Byrd-Jacobs, C. A., & Fuss, S. H. (2021). Diving into the streams and waves of constitutive and regenerative olfactory neurogenesis: insights from zebrafish. *Cell and Tissue Research*, *383*(1), 227–253. https://doi.org/10.1007/s00441-020-03334-2

Carleton, K. L., & Yourick, M. R. (2020). Axes of visual adaptation in the ecologically diverse family Cichlidae. *Seminars in Cell & Developmental Biology*, *106*, 43–52. https://doi.org/10.1016/j.semcdb.2020.04.015

Chang, J., Rabosky, D. L., Smith, S. A., & Alfaro, M. E. (2019). An r package and online resource for macroevolutionary studies using the ray-finned fish tree of life. *Methods in Ecology and Evolution / British Ecological Society*, *10*(7), 1118–1124.

https://doi.org/10.1111/2041-210X.13182

Chen, S., Zhou, Y., Chen, Y., & Gu, J. (2018). fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* , *34*(17), i884–i890. https://doi.org/10.1093/bioinformatics/bty560

Cong, X., Zheng, Q., Ren, W., Chéron, J.-B., Fiorucci, S., Wen, T., Zhang, C., Yu, H., Golebiowski, J., & Yu, Y. (2019). Zebrafish olfactory receptors ORAs differentially detect bile acids and bile salts. *The Journal of Biological Chemistry*, *294*(17), 6762–6771. https://doi.org/10.1074/jbc.RA118.006483

Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., Whitwham, A., Keane, T., McCarthy, S. A., Davies, R. M., & Li, H. (2021). Twelve years of SAMtools and BCFtools. *GigaScience*, *10*(2). https://doi.org/10.1093/gigascience/giab008

Darriba, D., Posada, D., Kozlov, A. M., Stamatakis, A., Morel, B., & Flouri, T. (2020). ModelTest-NG: A New and Scalable Tool for the Selection of DNA and Protein Evolutionary Models. *Molecular Biology and Evolution*, *37*(1), 291–294. https://doi.org/10.1093/molbev/msz189

Del Punta, K., Leinders-Zufall, T., Rodriguez, I., Jukam, D., Wysocki, C. J., Ogawa, S., Zufall, F., & Mombaerts, P. (2002). Deficient pheromone responses in mice lacking a cluster of vomeronasal receptor genes. *Nature*, *419*(6902), 70–74. https://doi.org/10.1038/nature00955

Edge, P., Bafna, V., & Bansal, V. (2017). HapCUT2: robust and accurate haplotype assembly for diverse sequencing technologies. *Genome Research*, *27*(5), 801–812. https://doi.org/10.1101/gr.213462.116

Escobar-Camacho, D., & Carleton, K. L. (2015). Sensory modalities in cichlid fish behavior. *Current Opinion in Behavioral Sciences*, *6*, 115–124.

https://doi.org/10.1016/j.cobeha.2015.11.002

Feulner, P. G. D., Schwarzer, J., Haesler, M. P., Meier, J. I., & Seehausen, O. (2018). A Dense Linkage Map of Lake Victoria Cichlids Improved the Pundamilia Genome Assembly and Revealed a Major QTL for Sex-Determination. *G3* , *8*(7), 2411–2420. https://doi.org/10.1534/g3.118.200207

Frei, D., De-Kayne, R., Selz, O. M., Seehausen, O., & Feulner, P. G. D. (2022). Genomic variation from an extinct species is retained in the extant radiation following speciation reversal. *Nature Ecology & Evolution*, *6*(4), 461–468. https://doi.org/10.1038/s41559-022-01665-7

Goddard, T. D., Huang, C. C., Meng, E. C., Pettersen, E. F., Couch, G. S., Morris, J. H., & Ferrin, T. E. (2018). UCSF ChimeraX: Meeting modern challenges in visualization and analysis. *Protein Science: A Publication of the Protein Society*, *27*(1), 14–25. https://doi.org/10.1002/pro.3235

Gopalakrishnan, S., Sinding, M.-H. S., Ramos-Madrigal, J., Niemann, J., Samaniego Castruita, J. A., Vieira, F. G., Carøe, C., Montero, M. de M., Kuderna, L., Serres, A., González-Basallote, V. M., Liu, Y.-H., Wang, G.-D., Marques-Bonet, T., Mirarab, S., Fernandes, C., Gaubert, P., Koepfli, K.-P., Budd, J., … Gilbert, M. T. P. (2018). Interspecific Gene Flow Shaped the Evolution of the Genus Canis. *Current Biology: CB*, *28*(21), 3441-3449.e5. https://doi.org/10.1016/j.cub.2018.08.041

Hallgren, J., Tsirigos, K. D., Pedersen, M. D., Armenteros, J. J. A., Marcatili, P., Nielsen, H., Krogh, A., & Winther, O. (2022). DeepTMHMM predicts alpha and beta transmembrane proteins using deep neural networks. In *bioRxiv* (p. 2022.04.08.487609). https://doi.org/10.1101/2022.04.08.487609

Irisarri, I., Singh, P., Koblmüller, S., Torres-Dowdall, J., Henning, F., Franchini, P., Fischer, C.,

Lemmon, A. R., Lemmon, E. M., Thallinger, G. G., Sturmbauer, C., & Meyer, A. (2018). Phylogenomics uncovers early hybridization and adaptive loci shaping the radiation of Lake Tanganyika cichlid fishes. *Nature Communications*, *9*(1), 3159. https://doi.org/10.1038/s41467-018-05479-9

Jiménez, J., Doerr, S., Martínez-Rosell, G., Rose, A. S., & De Fabritiis, G. (2017). DeepSite: protein-binding site predictor using 3D-convolutional neural networks. *Bioinformatics* , *33*(19), 3036–3042. https://doi.org/10.1093/bioinformatics/btx350

Johnson, T. C., Scholz, C. A., Talbot, M. R., Kelts, K., Ricketts, R. D., Ngobi, G., Beuning, K., Ssemmanda, I., I., & McGill, J. W. (1996). Late Pleistocene desiccation of lake Victoria and rapid evolution of cichlid fishes. *Science (New York, N.Y.)*, *273*(5278), 1091–1093. https://doi.org/10.1126/science.273.5278.1091

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., … Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, *596*(7873), 583–589. https://doi.org/10.1038/s41586-021-03819-2

Kato, S., Arakaki, S., Nagano, A. J., Kikuchi, K., & Hirase, S. (2023). Genomic landscape of introgression from the ghost lineage in a gobiid fish uncovers the generality of forces shaping hybrid genomes. *Molecular Ecology*. https://doi.org/10.1111/mec.17216

Katoh, K., & Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution*, *30*(4), 772–780. https://doi.org/10.1093/molbev/mst010

Kawamura, R., & Nikaido, M. (2022). c-fos expression in the olfactory epithelium of the East African cichlid (Haplochromis chilotes) in response to odorant exposure. *The Journal*

of Experimental Biology. https://doi.org/10.1242/jeb.244737

Keller-Costa, T., Canário, A. V. M., & Hubbard, P. C. (2015). Chemical communication in cichlids: A mini-review. *General and Comparative Endocrinology*, *221*, 64–74. https://doi.org/10.1016/j.ygcen.2015.01.001

Keller-Costa, T., Hubbard, P. C., Paetz, C., Nakamura, Y., da Silva, J. P., Rato, A., Barata, E. N., Schneider, B., & Canario, A. V. M. (2014). Identity of a tilapia pheromone released by dominant males that primes females for reproduction. *Current Biology: CB*, *24*(18), 2130–2135. https://doi.org/10.1016/j.cub.2014.07.049

Kocher, T. D. (2004). Adaptive evolution and explosive speciation: the cichlid fish model. *Nature Reviews. Genetics*, *5*(4), 288–298. https://doi.org/10.1038/nrg1316

Kochnev, Y., Hellemann, E., Cassidy, K. C., & Durrant, J. D. (2020). Webina: an open-source library and web app that runs AutoDock Vina entirely in the web browser. *Bioinformatics* , *36*(16), 4513–4515. https://doi.org/10.1093/bioinformatics/btaa579

Kozlov, A. M., Darriba, D., Flouri, T., Morel, B., & Stamatakis, A. (2019). RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. *Bioinformatics* , *35*(21), 4453–4455. https://doi.org/10.1093/bioinformatics/btz305

Larsson, A. (2014). AliView: a fast and lightweight alignment viewer and editor for large datasets. *Bioinformatics (Oxford, England)*, *30*(22), 3276–3278. https://doi.org/10.1093/bioinformatics/btu531

Letunic, I., & Bork, P. (2024). Interactive Tree of Life (iTOL) v6: recent updates to the phylogenetic tree display and annotation tool. *Nucleic Acids Research*, *52*(W1), W78–W82. https://doi.org/10.1093/nar/gkae268

Loh, Y.-H. E., Bezault, E., Muenzel, F. M., Roberts, R. B., Swofford, R., Barluenga, M., Kidd, C. E., Howe, A. E., Di Palma, F., Lindblad-Toh, K., Hey, J., Seehausen, O., Salzburger,

W., Kocher, T. D., & Streelman, J. T. (2013). Origins of shared genetic variation in African cichlids. *Molecular Biology and Evolution*, *30*(4), 906–917. https://doi.org/10.1093/molbev/mss326

Malinsky, M., Svardal, H., Tyers, A. M., Miska, E. A., Genner, M. J., Turner, G. F., & Durbin, R. (2018). Whole-genome sequences of Malawi cichlids reveal multiple radiations interconnected by gene flow. *Nature Ecology & Evolution*, *2*(12), 1940–1955. https://doi.org/10.1038/s41559-018-0717-x

Matschiner, M., Böhne, A., Ronco, F., & Salzburger, W. (2020). The genomic timeline of cichlid fish diversification across continents. *Nature Communications*, *11*(1), 5895. https://doi.org/10.1038/s41467-020-17827-9

McGee, M. D., Borstein, S. R., Meier, J. I., Marques, D. A., Mwaiko, S., Taabu, A., Kishe, M. A., O'Meara, B., Bruggmann, R., Excoffier, L., & Seehausen, O. (2020). The ecological and genomic basis of explosive adaptive radiation. *Nature*, *586*(7827), 75–79. https://doi.org/10.1038/s41586-020-2652-7

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., & DePristo, M. A. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, *20*(9), 1297–1303. https://doi.org/10.1101/gr.107524.110

Meier, J. I., Marques, D. A., Mwaiko, S., Wagner, C. E., Excoffier, L., & Seehausen, O. (2017). Ancient hybridization fuels rapid cichlid fish adaptive radiations. *Nature Communications*, *8*, 14363. https://doi.org/10.1038/ncomms14363

Nakamura, H., Aibara, M., Kajitani, R., Mrosso, H. D. J., Mzighani, S. I., Toyoda, A., Itoh, T., Okada, N., & Nikaido, M. (2021). Genomic Signatures for Species-Specific Adaptation in Lake Victoria Cichlids Derived from Large-Scale Standing Genetic

Variation. *Molecular Biology and Evolution*, *38*(8), 3111–3125. https://doi.org/10.1093/molbev/msab084

Nakamura, H., Aibara, M., & Nikaido, M. (2023). Ancient standing genetic variation facilitated the adaptive radiation of Lake Victoria cichlids. *Genes & Genetic Systems*. https://doi.org/10.1266/ggs.23-00024

Nicoli, A., Dunkel, A., Giorgino, T., de Graaf, C., & Di Pizio, A. (2022). Classification Model for the Second Extracellular Loop of Class A GPCRs. *Journal of Chemical Information and Modeling*, *62*(3), 511–522. https://doi.org/10.1021/acs.jcim.1c01056

Nikaido, M., Ota, T., Hirata, T., Suzuki, H., Satta, Y., Aibara, M., Mzighani, S. I., Sturmbauer, C., Hagino-Yamagishi, K., & Okada, N. (2014). Multiple episodic evolution events in V1R receptor genes of East-African cichlids. *Genome Biology and Evolution*, *6*(5), 1135–1144. https://doi.org/10.1093/gbe/evu086

Ota, T., Nikaido, M., Suzuki, H., Hagino-Yamagishi, K., & Okada, N. (2012). Characterization of V1R receptor (ora) genes in Lake Victoria cichlids. *Gene*, *499*(2), 273–279. https://doi.org/10.1016/j.gene.2012.03.002

Palkopoulou, E., Lipson, M., Mallick, S., Nielsen, S., Rohland, N., Baleka, S., Karpinski, E., Ivancevic, A. M., To, T.-H., Kortschak, R. D., Raison, J. M., Qu, Z., Chin, T.-J., Alt, K. W., Claesson, S., Dalén, L., MacPhee, R. D. E., Meller, H., Roca, A. L., … Reich, D. (2018). A comprehensive genomic history of extinct and living elephants. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(11), E2566–E2574. https://doi.org/10.1073/pnas.1720554115

Pfister, P., & Rodriguez, I. (2005). Olfactory expression of a single and highly variable V1r pheromone receptor-like gene in fish species. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(15), 5489–5494.

https://doi.org/10.1073/pnas.0402581102

Plenderleith, M., van Oosterhout, C., Robinson, R. L., & Turner, G. F. (2005). Female preference for conspecific males based on olfactory cues in a Lake Malawi cichlid fish. *Biology Letters*, *1*(4), 411–414. https://doi.org/10.1098/rsbl.2005.0355

Policarpo, M., Bemis, K. E., Laurenti, P., Legendre, L., Sandoz, J.-C., Rétaux, S., & Casane, D. (2022). Coevolution of the olfactory organ and its receptor repertoire in ray-finned fishes. *BMC Biology*, *20*(1), 195. https://doi.org/10.1186/s12915-022-01397-x

Rabosky, D. L., Chang, J., Title, P. O., Cowman, P. F., Sallan, L., Friedman, M., Kaschner, K., Garilao, C., Near, T. J., Coll, M., & Alfaro, M. E. (2018). An inverse latitudinal gradient in speciation rate for marine fishes. *Nature*, *559*(7714), 392–395. https://doi.org/10.1038/s41586-018-0273-1

Rometsch, S. J., Torres-Dowdall, J., & Meyer, A. (2020). Evolutionary dynamics of pre- and postzygotic reproductive isolation in cichlid fishes. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *375*(1806), 20190535. https://doi.org/10.1098/rstb.2019.0535

Ronco, F., Matschiner, M., Böhne, A., Boila, A., Büscher, H. H., El Taher, A., Indermaur, A., Malinsky, M., Ricci, V., Kahmen, A., Jentoft, S., & Salzburger, W. (2021). Drivers and dynamics of a massive adaptive radiation in cichlid fishes. *Nature*, *589*(7840), 76–81. https://doi.org/10.1038/s41586-020-2930-4

Salzburger, W. (2018). Understanding explosive diversification through cichlid fish genomics. *Nature Reviews. Genetics*, *19*(11), 705–717. https://doi.org/10.1038/s41576-018-0043-9

Salzburger, W., Van Bocxlaer, B., & Cohen, A. S. (2014). Ecology and Evolution of the African Great Lakes and Their Faunas. *Annual Review of Ecology, Evolution, and*

Systematics, *45*(1), 519–545. https://doi.org/10.1146/annurev-ecolsys-120213-091804

Santos, M. E., Lopes, J. F., & Kratochwil, C. F. (2023). East African cichlid fishes. *EvoDevo*, *14*(1), 1. https://doi.org/10.1186/s13227-022-00205-5

Saraiva, L. R., & Korsching, S. I. (2007). A novel olfactory receptor gene family in teleost fish. *Genome Research*, *17*(10), 1448–1457. https://doi.org/10.1101/gr.6553207

Seehausen, O., Terai, Y., Magalhaes, I. S., Carleton, K. L., Mrosso, H. D. J., Miyagi, R., van der Sluijs, I., Schneider, M. V., Maan, M. E., Tachida, H., Imai, H., & Okada, N. (2008). Speciation through sensory drive in cichlid fish. *Nature*, *455*(7213), 620–626. https://doi.org/10.1038/nature07285

Seehausen, O., & Wagner, C. E. (2014). Speciation in Freshwater Fishes. *Annual Review of Ecology, Evolution, and Systematics*, *45*(1), 621–651. https://doi.org/10.1146/annurev-ecolsys-120213-091818

Suyama, M., Torrents, D., & Bork, P. (2006). PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Research*, *34*(Web Server issue), W609-12. https://doi.org/10.1093/nar/gkl315

Svardal, H., Salzburger, W., & Malinsky, M. (2021). Genetic Variation and Hybridization in Evolutionary Radiations of Cichlid Fishes. *Annual Review of Animal Biosciences*, *9*, 55–79. https://doi.org/10.1146/annurev-animal-061220-023129

Terai, Y., Seehausen, O., Sasaki, T., Takahashi, K., Mizoiri, S., Sugawara, T., Sato, T., Watanabe, M., Konijnendijk, N., Mrosso, H. D. J., Tachida, H., Imai, H., Shichida, Y., & Okada, N. (2006). Divergent selection on opsins drives incipient speciation in Lake Victoria cichlids. *PLoS Biology*, *4*(12), e433. https://doi.org/10.1371/journal.pbio.0040433

Torrens-Fontanals, M., Tourlas, P., Doerr, S., & De Fabritiis, G. (2024). PlayMolecule Viewer: A toolkit for the visualization of molecules and other data. *Journal of Chemical Information and Modeling*, *64*(3), 584–589. https://doi.org/10.1021/acs.jcim.3c01776

Tumescheit, C., Firth, A. E., & Brown, K. (2022). CIAlign: A highly customisable command line tool to clean, interpret and visualise multiple sequence alignments. *PeerJ*, *10*, e12983. https://doi.org/10.7717/peerj.12983

Vasimuddin, M., Misra, S., Li, H., & Aluru, S. (2019). Efficient Architecture-Aware Acceleration of BWA-MEM for Multicore Systems. *2019 IEEE International Parallel and Distributed Processing Symposium (IPDPS)*, 314–324. https://doi.org/10.1109/IPDPS.2019.00041

Verzijden, M. N., & ten Cate, C. (2007). Early learning influences species assortative mating preferences in Lake Victoria cichlid fish. *Biology Letters*, *3*(2), 134–136. https://doi.org/10.1098/rsbl.2006.0601

Villanueva, R. A. M., & Chen, Z. J. (2019). ggplot2: Elegant Graphics for Data Analysis (2nd ed.). *Measurement: Interdisciplinary Research and Perspectives*, *17*(3), 160–167. https://doi.org/10.1080/15366367.2019.1565254

Wang, L.-G., Lam, T. T.-Y., Xu, S., Dai, Z., Zhou, L., Feng, T., Guo, P., Dunn, C. W., Jones, B. R., Bradley, T., Zhu, H., Guan, Y., Jiang, Y., & Yu, G. (2020). Treeio: An R package for phylogenetic tree input and output with richly annotated and associated data. *Molecular Biology and Evolution*, *37*(2), 599–603. https://doi.org/10.1093/molbev/msz240

Weiss, J. D., Cotterill, F. P. D., & Schliewen, U. K. (2015). Lake Tanganyika--a “melting pot” of ancient and young cichlid lineages (Teleostei: Cichlidae)? *PloS One*, *10*(4), e0125043. https://doi.org/10.1371/journal.pone.0125043

Wickham H, François R, Henry L, Müller K, Vaughan D. (2023). *dplyr: A Grammar of Data Manipulation* (1.1.4) [Computer software]. https://github.com/tidyverse/dplyr, https://dplyr.tidyverse.org

Wickham H, Hester J, Bryan J. (2024). *readr: Read Rectangular Text Data* (2.1.5) [Computer software]. https://github.com/tidyverse/readr, https://readr.tidyverse.org

Xu, W., Wu, L., Liu, S., Liu, X., Cao, X., Zhou, C., Zhang, J., Fu, Y., Guo, Y., Wu, Y., Tan, Q., Wang, L., Liu, J., Jiang, L., Fan, Z., Pei, Y., Yu, J., Cheng, J., Zhao, S., … Hua, T. (2022). Structural basis for strychnine activation of human bitter taste receptor TAS2R46. *Science*, *377*(6612), 1298–1304. https://doi.org/10.1126/science.abo1633

Yang, Z. (2007). PAML 4: phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution*, *24*(8), 1586–1591. https://doi.org/10.1093/molbev/msm088

Yu, G., Smith, D. K., Zhu, H., Guan, Y., & Lam, T. T.-Y. (2017). Ggtree: An r package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods in Ecology and Evolution*, *8*(1), 28–36. https://doi.org/10.1111/2041-210x.12628

Zapilko, V., & Korsching, S. I. (2016). Tetrapod V1R-like ora genes in an early-diverging ray-finned fish species: the canonical six ora gene repertoire of teleost fish resulted from gene loss in a larger ancestral repertoire. *BMC Genomics*, *17*, 83. https://doi.org/10.1186/s12864-016-2399-6

Zhang, D., Tang, L., Cheng, Y., Hao, Y., Xiong, Y., Song, G., Qu, Y., Rheindt, F. E., Alström, P., Jia, C., & Lei, F. (2019). “Ghost Introgression” As a Cause of Deep Mitochondrial Divergence in a Bird Species Complex. *Molecular Biology and Evolution*, *36*(11), 2375–2386. https://doi.org/10.1093/molbev/msz170

Zhou, L., Feng, T., Xu, S., Gao, F., Lam, T. T., Wang, Q., Wu, T., Huang, H., Zhan, L., Li, L.,

Guan, Y., Dai, Z., & Yu, G. (2022). Ggmsa: A visual exploration tool for multiple sequence alignment and associated data. *Briefings in Bioinformatics*, *23*(4). https://doi.org/10.1093/bib/bbac222

Author Contributions

M.N. supervised the project. S.T. conducted all bioinformatics and evolutionary analyses. S.T. and M.N. wrote the manuscript. Z.Z., M.A., and T.N. provided their expertise and commented on the paper. All authors read and approved the final manuscript.

Tables and Figures

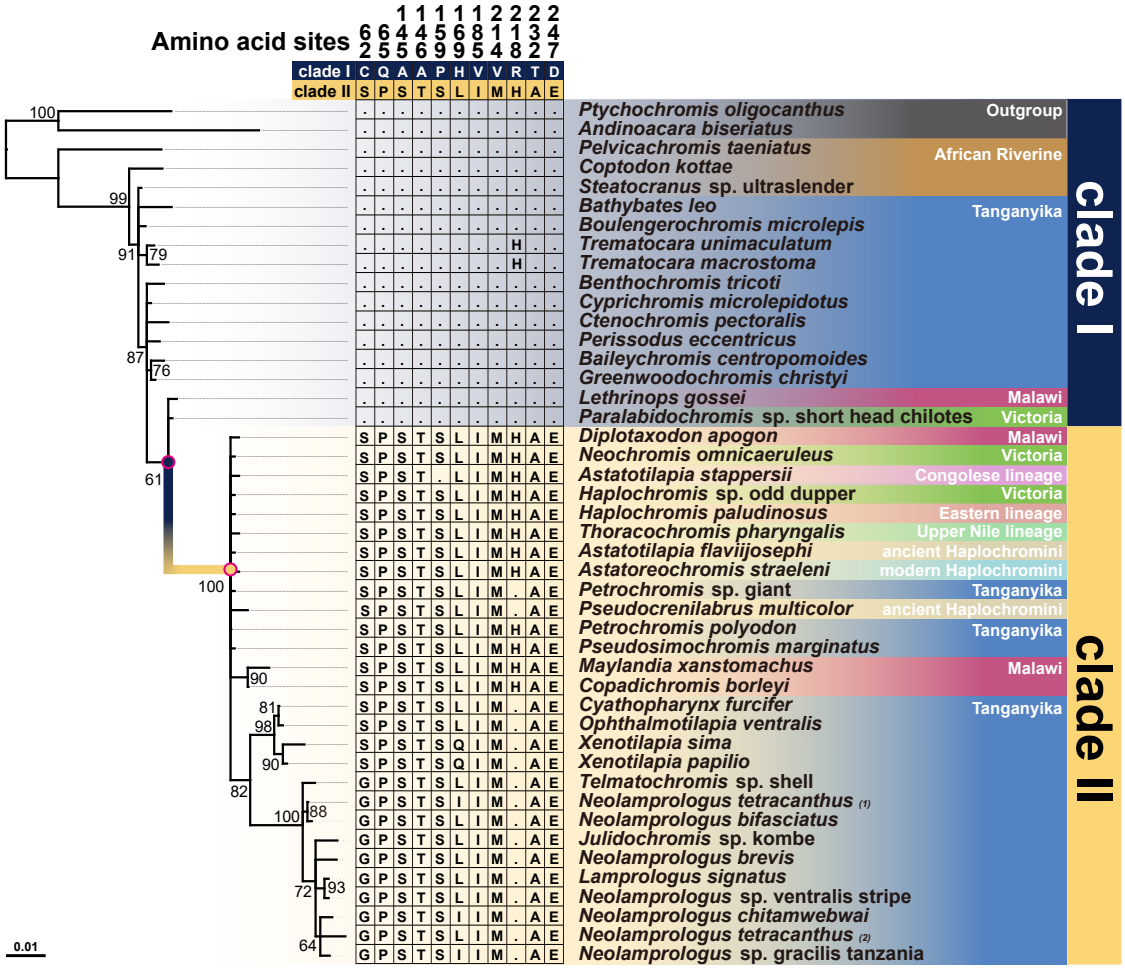


Figure 1 A phylogenetic tree and sequence alignment of the V1R2 gene in cichlids. Bootstrap values are shown only for nodes with support values of 60 or higher. The sequences are divided into two highly divergent alleles (clade I and clade II). Only diagnostic amino acid sites between clade I and clade II were shown. Dots indicate identity with the top sequence (clade I). The phase of each heterozygous allele in the individuals is indicated in parentheses. Observed recombination-derived sequences, which are exceptionally rare, have been excluded. Note that sequences from each lake (Tanganyika, Malawi, Victoria) do not form monophyletic groups.

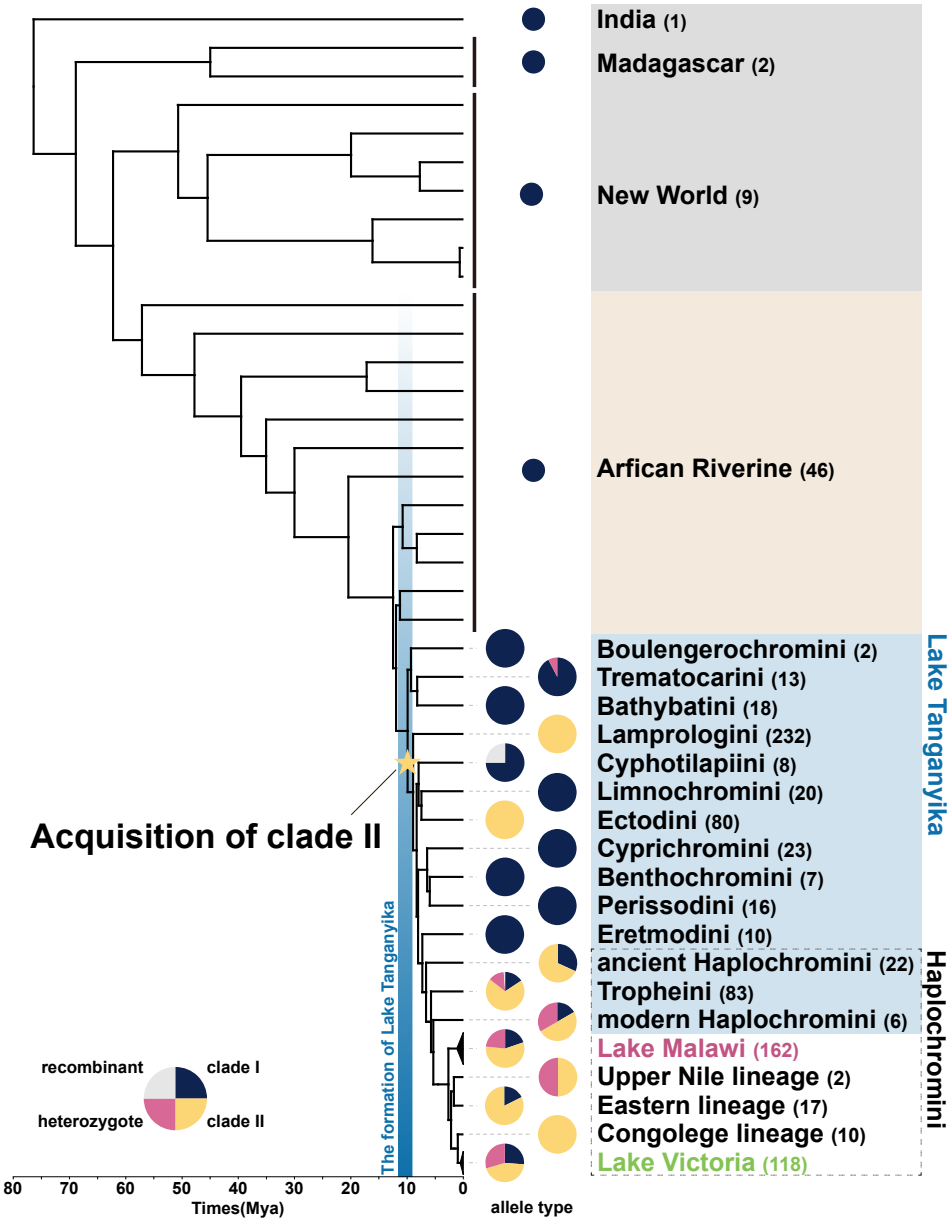


Figure 2 Frequencies of each allele of the V1R2 gene in each cichlid lineage. We calculated the frequencies of clade I and clade II alleles in each lineage. Clade I are shown in dark blue, clade II in yellow, heterozygotes in pink, and recombinants in grey. The phylogenetic tree and divergence times are based on previous studies (Matschiner et al., 2020; Meier et al., 2017; Nakamura et al., 2023; Ronco et al., 2021). Numbers in parentheses next to lineage names show the number of samples. A star represents the estimated times of acquisition of clade II.

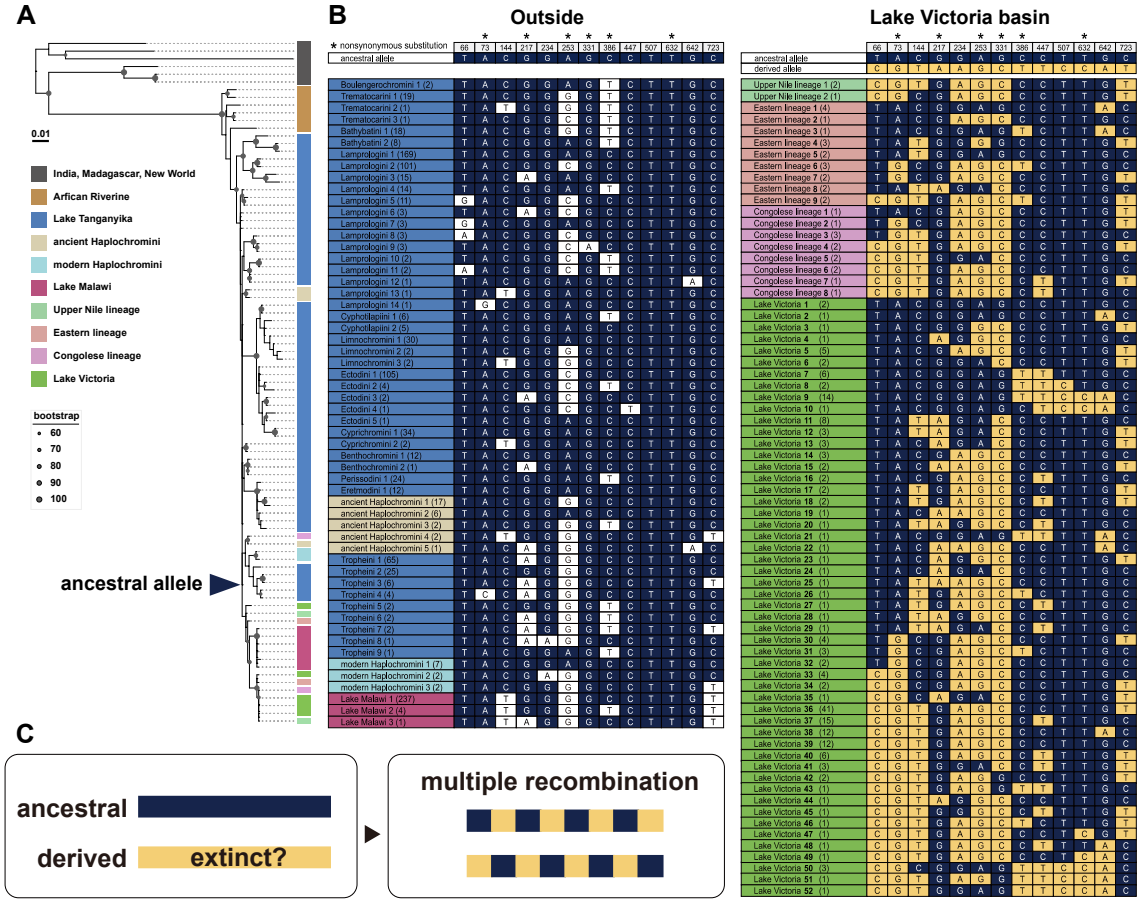


Figure 3 Recombination events observed in the V1R6 gene of cichlids. (A) Phylogenetic tree of the V1R6 gene. Bootstrap values are shown only for nodes with support values of 60 or higher. (B) All haplotypes of the V1R6 gene. Only sites that are polymorphic in Lake Victoria lineage are displayed. Numbers in parentheses following lineage names show the number of samples. Asterisks above site numbers denote nonsynonymous substitutions. Ancestral and derived alleles at each site were defined based on the frequency of nucleotides across the sequences. Sites matching the ancestral allele are highlighted in dark blue, and those matching the derived allele are highlighted in yellow. (C) Schematic diagram of recombination.

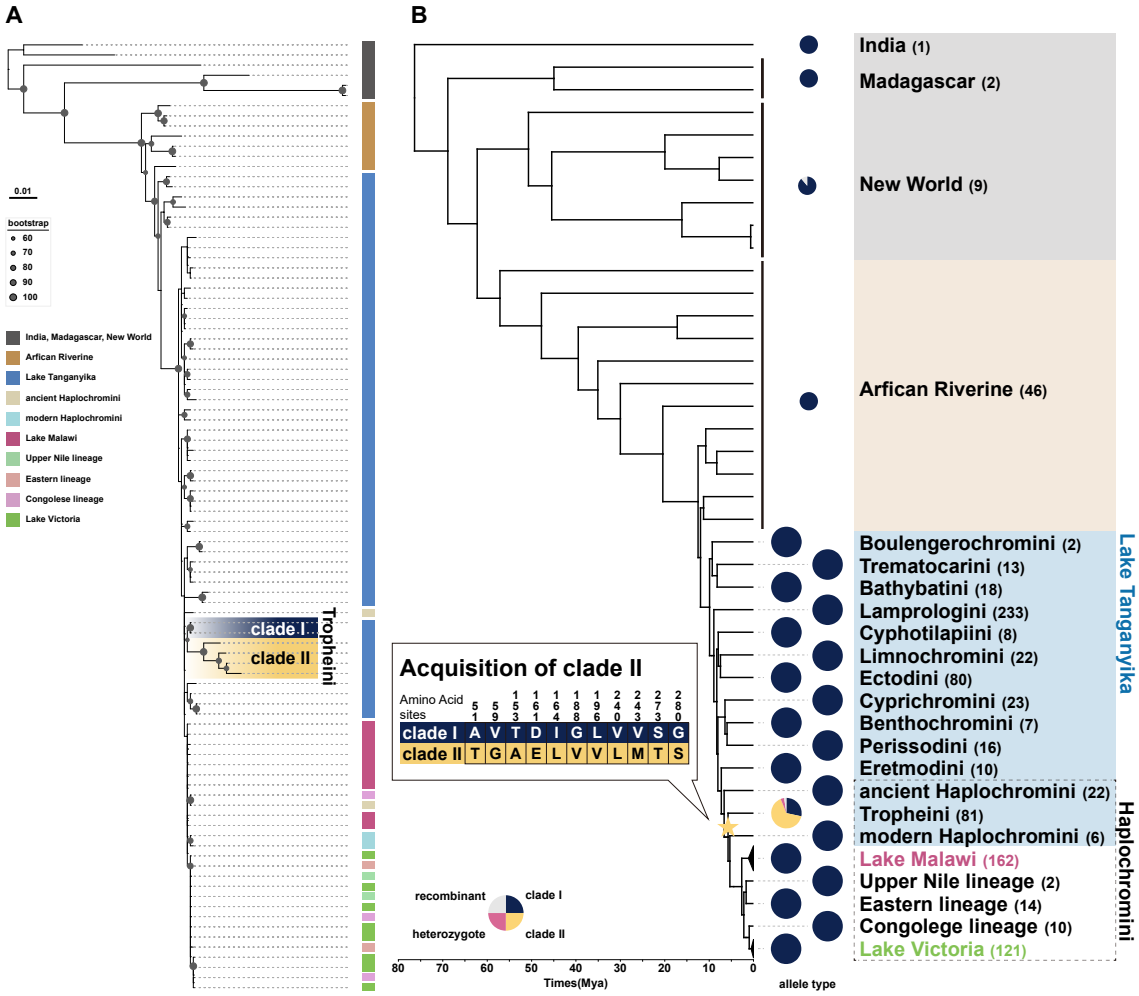


Figure 4 Two highly divergent alleles observed in the V1R1 gene of Tropheini. (A) A phylogenetic tree of the V1R1 gene. Bootstrap values are shown only for nodes with support values of 60 or higher. (B) Frequencies of each allele of the V1R1 gene in each cichlid lineage. Clade I are shown in dark blue, clade II in yellow, heterozygotes in pink, and recombinants in grey. The phylogenetic tree and divergence times are based on previous studies (Matschiner et al., 2020; Meier et al., 2017; Nakamura et al., 2023; Ronco et al., 2021). Numbers in parentheses next to the lineage names show the number of samples. A star represents the estimated timing of acquisition of clade II.

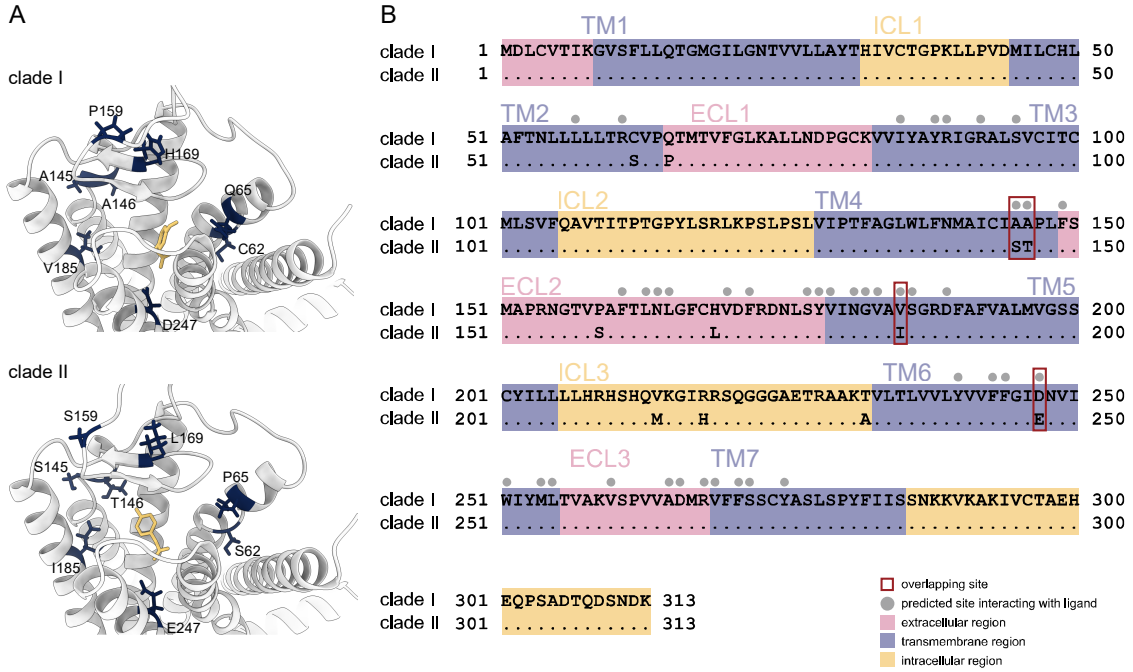


Figure 5 3D structure of cichlid V1R2. (A) 3D structure of V1R2 viewed from the extracellular side. Amino acids that differ between Clade I and Clade II are shown in dark blue. The ligand used for analysis (4HPAA) is shown in yellow. (B) Predicted transmembrane regions of cichlid V1R2. The transmembrane regions were predicted using the sequence of Clade I. Dots indicate amino acids identical to those in Clade I. The extracellular regions are shown in pink, transmembrane regions in blue, and intracellular regions in yellow. TM: transmembrane; ECL: extracellular loop; ICL: intracellular loop. Gray dots indicate predicted ligand-binding sites. Red squares indicate ligand-binding sites that are also sites of substitution between alleles. Mutations were found to occur at ligand-binding sites significantly more often than expected by chance (Fisher's exact test: $p = 0.0389$).

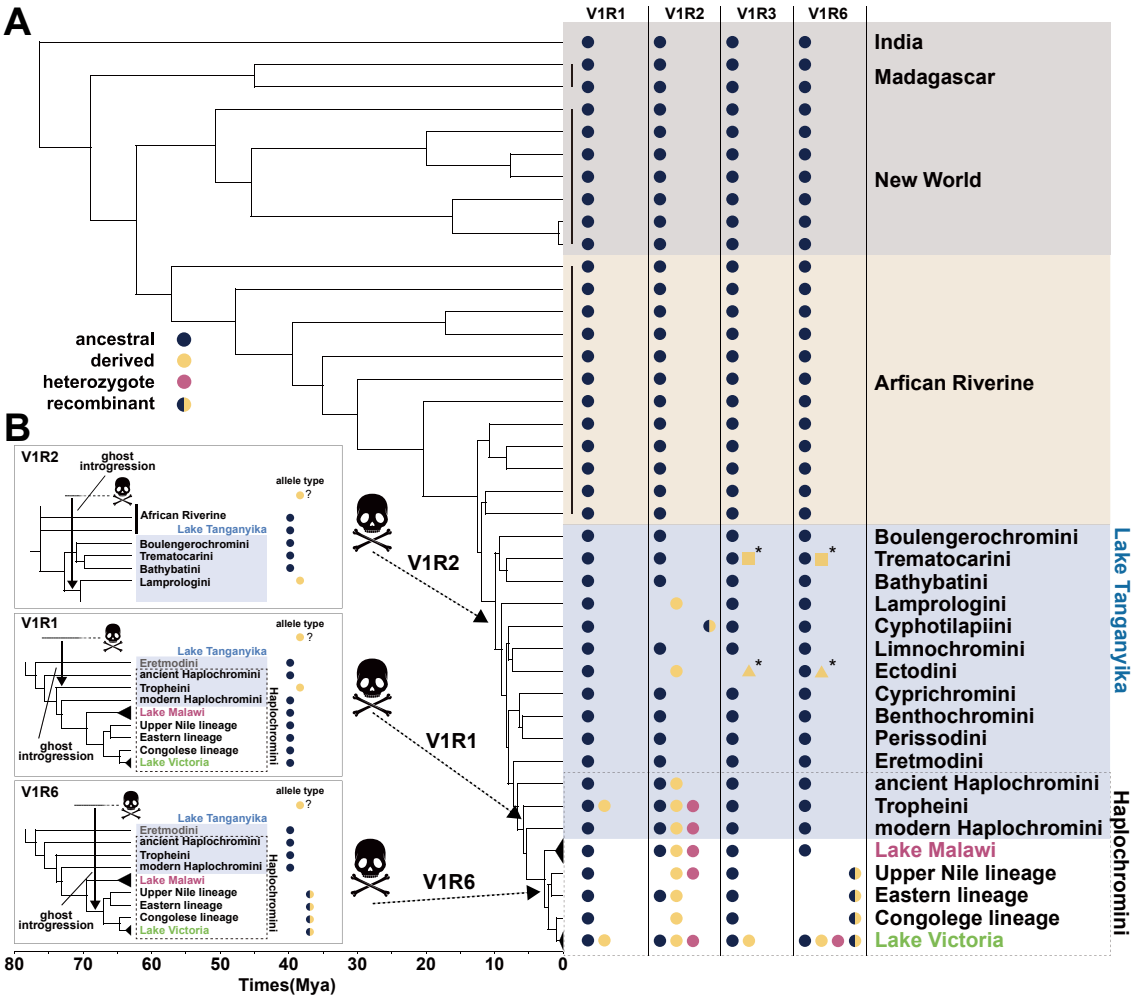


Figure 6 Possible involvement of ghost introgressions in cichlids (A) Sequence diversity of V1R genes in cichlids. Allele types of V1R genes are represented by colored dots for each lineage. The phylogenetic tree and divergence times are based on previous studies (Matschiner et al., 2020; Meier et al., 2017; Nakamura et al., 2023; Ronco et al., 2021). Note that the derived allele (clade II) of V1R2 is shared across lineages, whereas the derived sequences of V1R3 and V1R6 differ among lineages (asterisks). (B) Predicted mechanisms of allele acquisition for V1R1, V1R2, and V1R6 genes. Arrows indicate the most parsimonious estimated times of introgression.

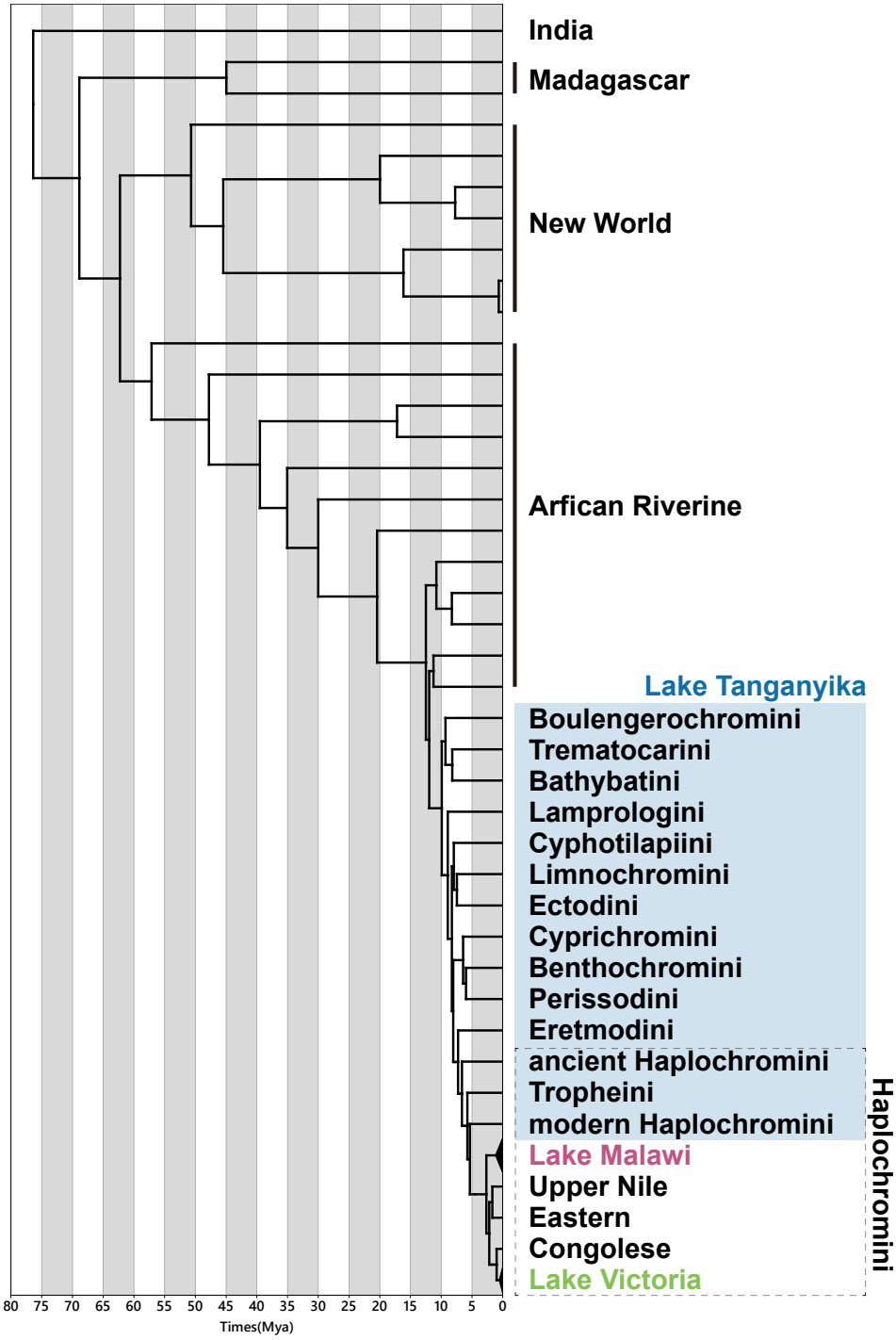


Figure S1 The lineage names in this study were defined with reference to previous studies (Matschiner et al., 2020; Meier et al., 2017; Nakamura et al., 2023; Ronco et al., 2021). Haplochromini species that diverged before Tropheini are defined as "ancient Haplochromini." Divergence times are expressed in millions of years (Mya).

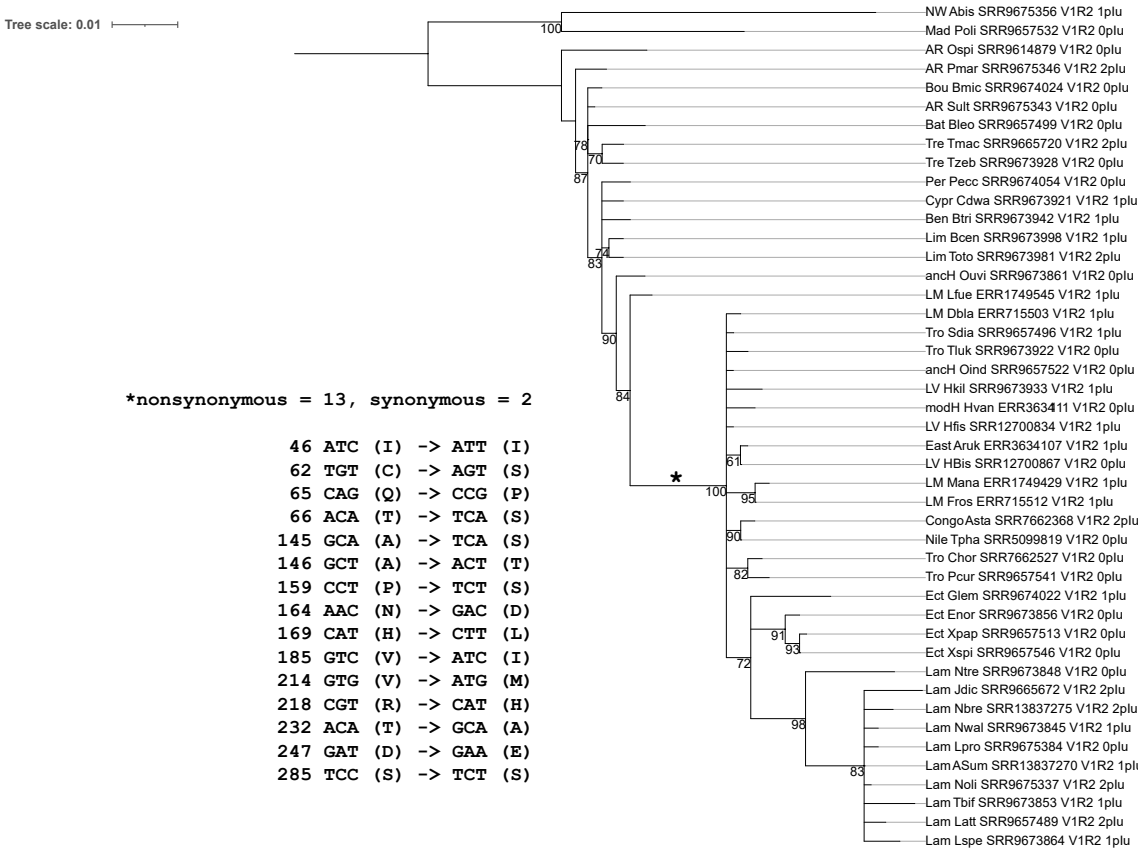


Figure S2 A phylogenetic tree and sequence alignment of the V1R2 gene in cichlids (Dataset 1). The phylogenetic tree displays bootstrap values only for nodes with support values of 60 or higher. The left shows nonsynonymous and synonymous substitution sites at positions marked with an asterisk.

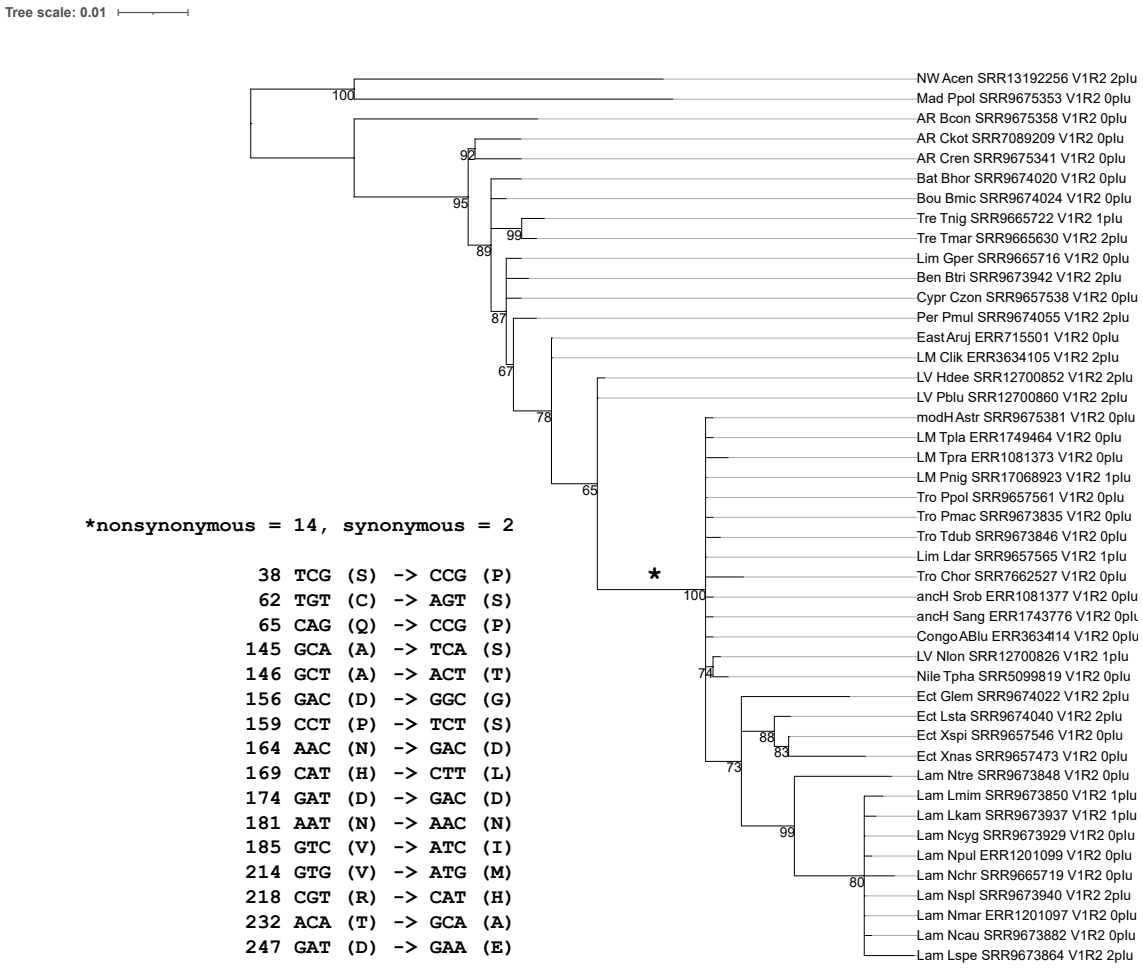


Figure S3 A phylogenetic tree and sequence alignment of the V1R2 gene in cichlids (Dataset 2). The phylogenetic tree displays bootstrap values only for nodes with support values of 60 or higher. The left shows nonsynonymous and synonymous substitution sites at positions marked with an asterisk.

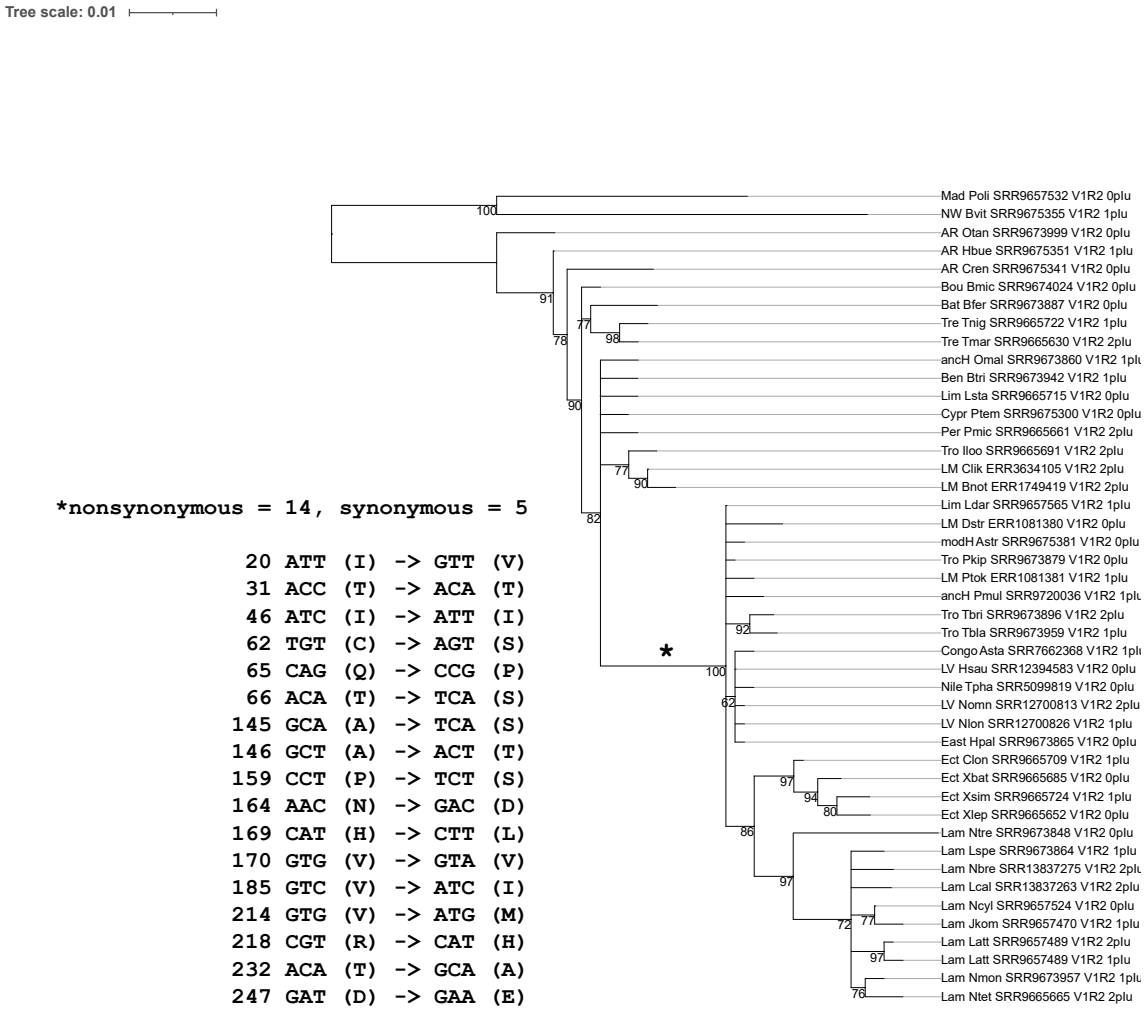


Figure S4 A phylogenetic tree and sequence alignment of the V1R2 gene in cichlids (Dataset 3). The phylogenetic tree displays bootstrap values only for nodes with support values of 60 or higher. The left shows nonsynonymous and synonymous substitution sites at positions marked with an asterisk.

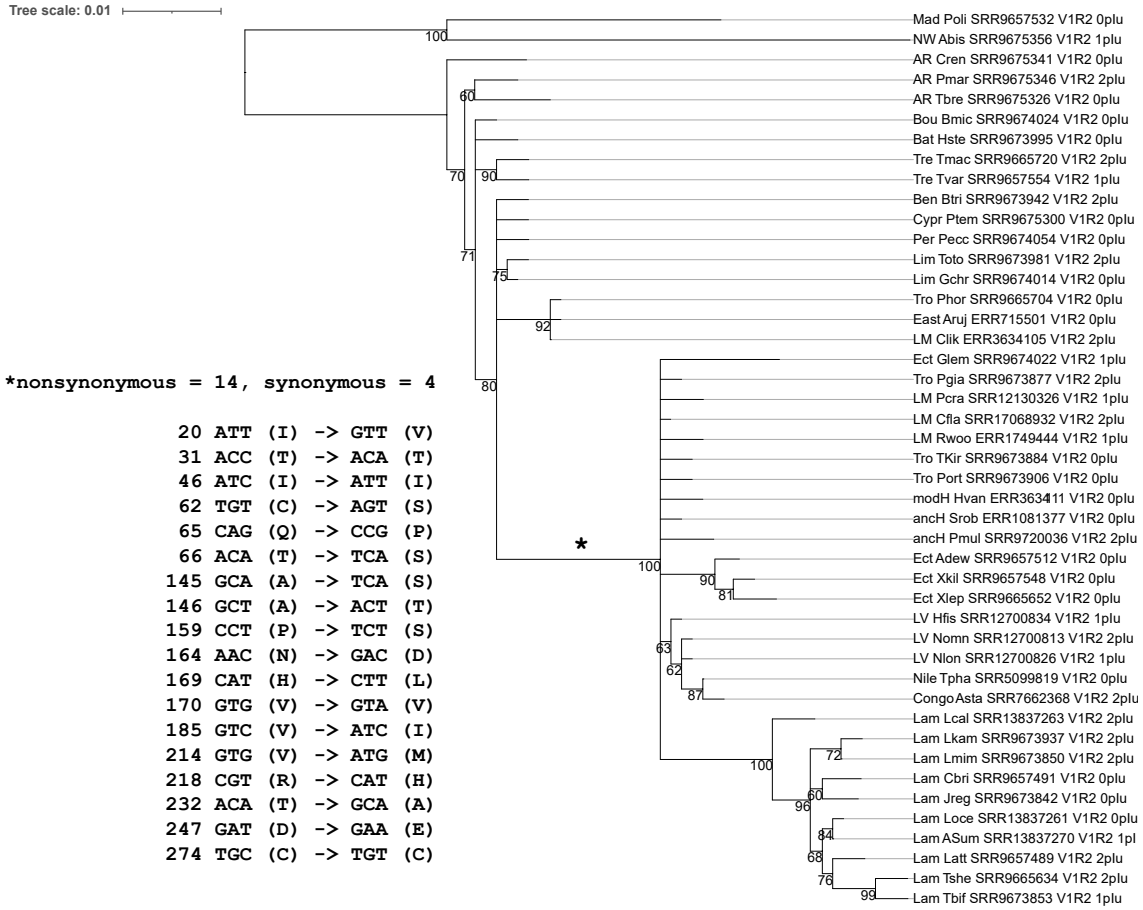


Figure S5 A phylogenetic tree and sequence alignment of the V1R2 gene in cichlids (Dataset 4). The phylogenetic tree displays bootstrap values only for nodes with support values of 60 or higher. The left shows nonsynonymous and synonymous substitution sites at positions marked with an asterisk.

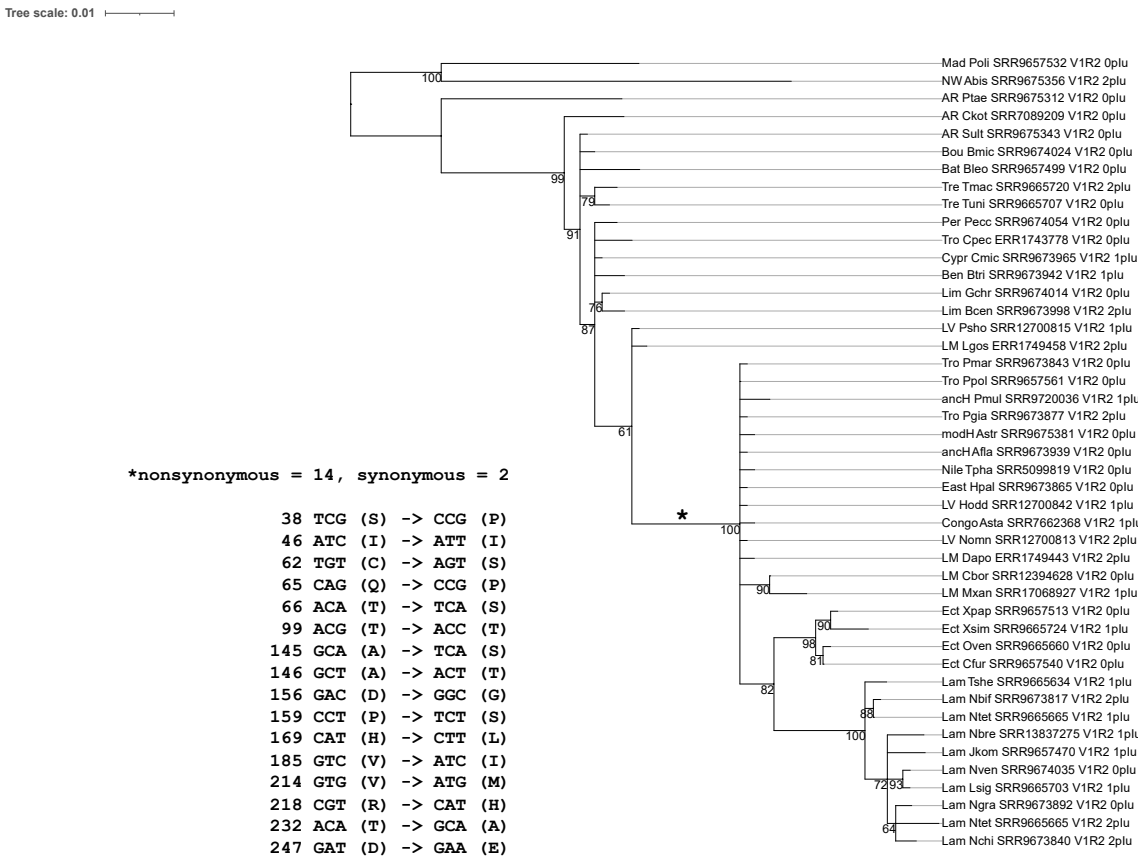


Figure S6 A phylogenetic tree and sequence alignment of the V1R2 gene in cichlids (Dataset 5). The phylogenetic tree displays bootstrap values only for nodes with support values of 60 or higher. The left shows nonsynonymous and synonymous substitution sites at positions marked with an asterisk.

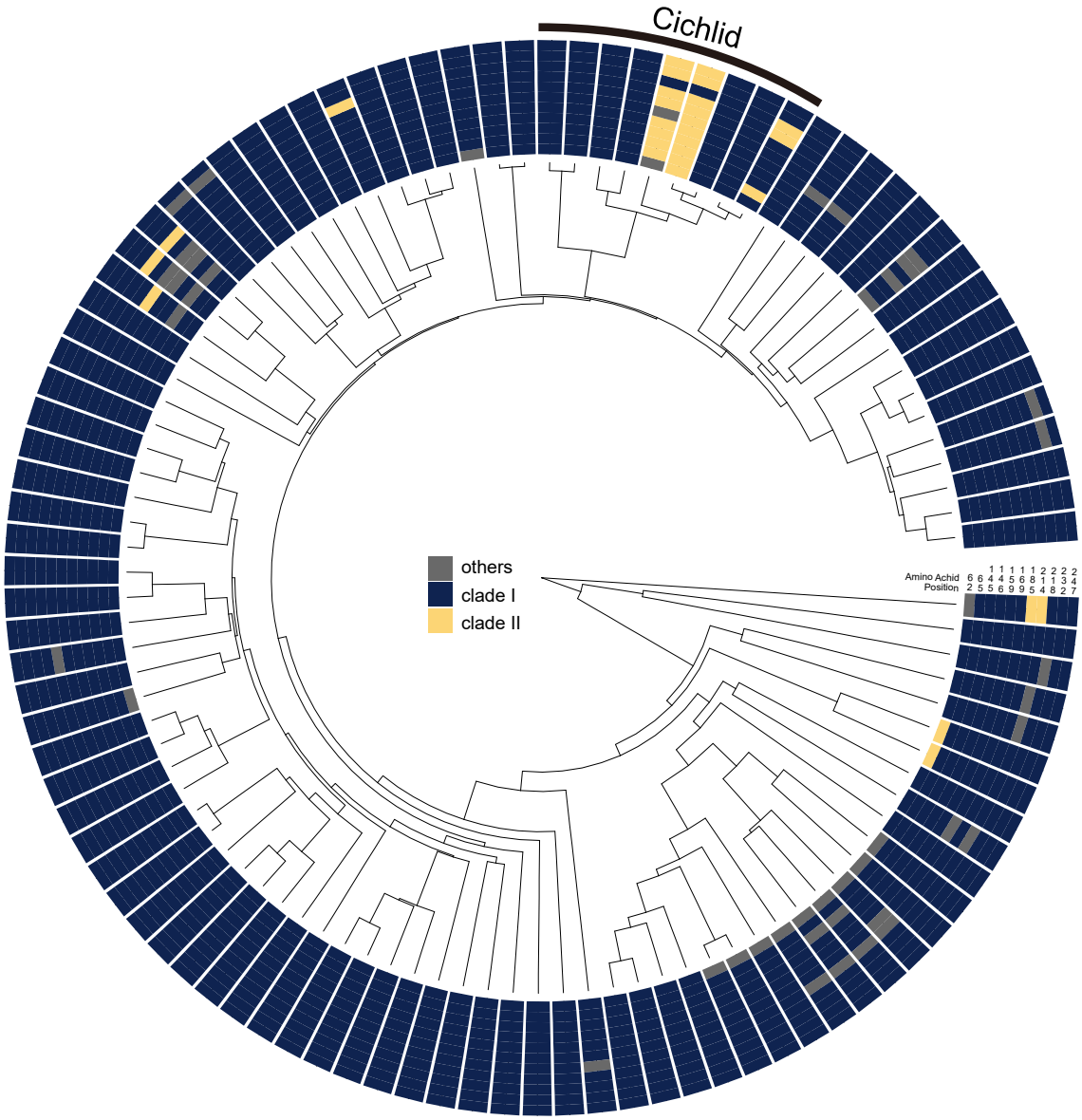


Figure S7 V1R2 alignment of 102 ray-finned fish species. The alignment displays sites corresponding to the two alleles (clade I and clade II). Sites matching clade I are highlighted in dark blue, those matching clade II are highlighted in yellow, and all other sites are shown in grey.

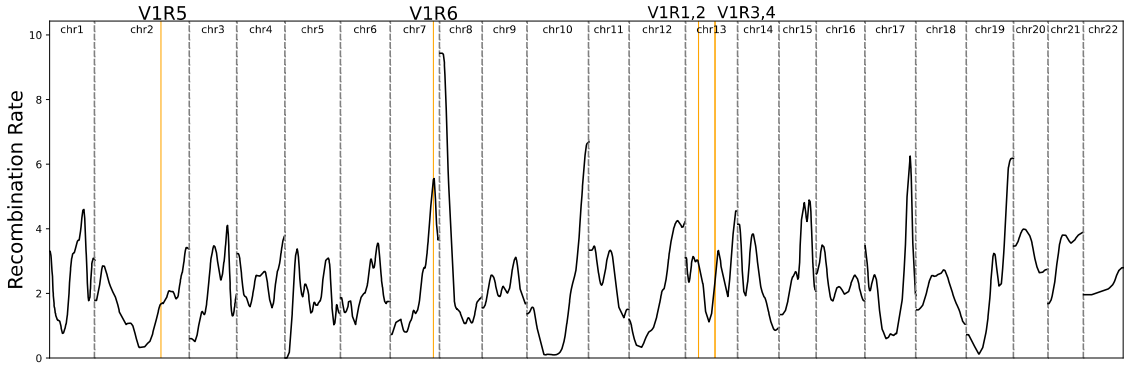


Figure S8 Recombination rate (cM/Mb) per chromosome in *Pundamilia nyererei* from the Lake Victoria lineage (Feulner et al., 2018). The positions of V1R genes are indicated by yellow lines.

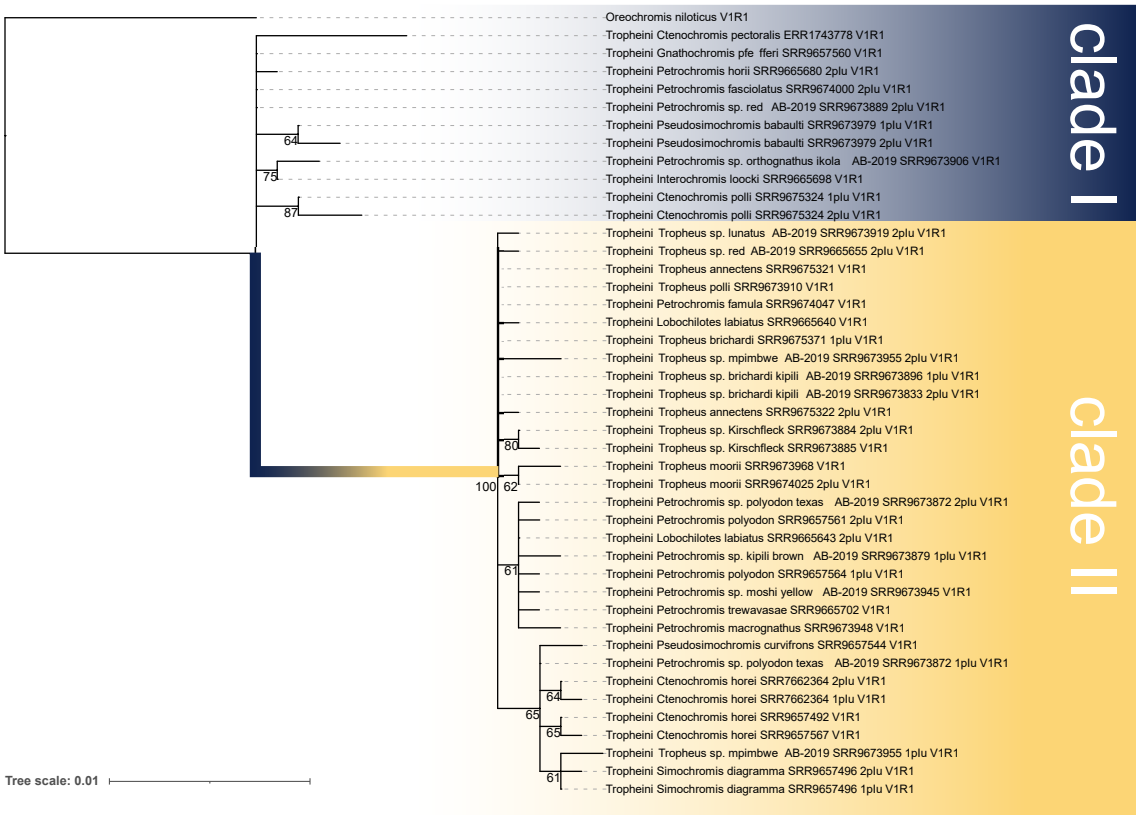


Figure S9 A phylogenetic tree of the V1R1 gene in Tropheini. Bootstrap values are shown only for nodes with support values of 60 or higher. The sequences are divided into two highly divergent alleles (clade I and clade II). Recombination-derived sequences have been excluded.

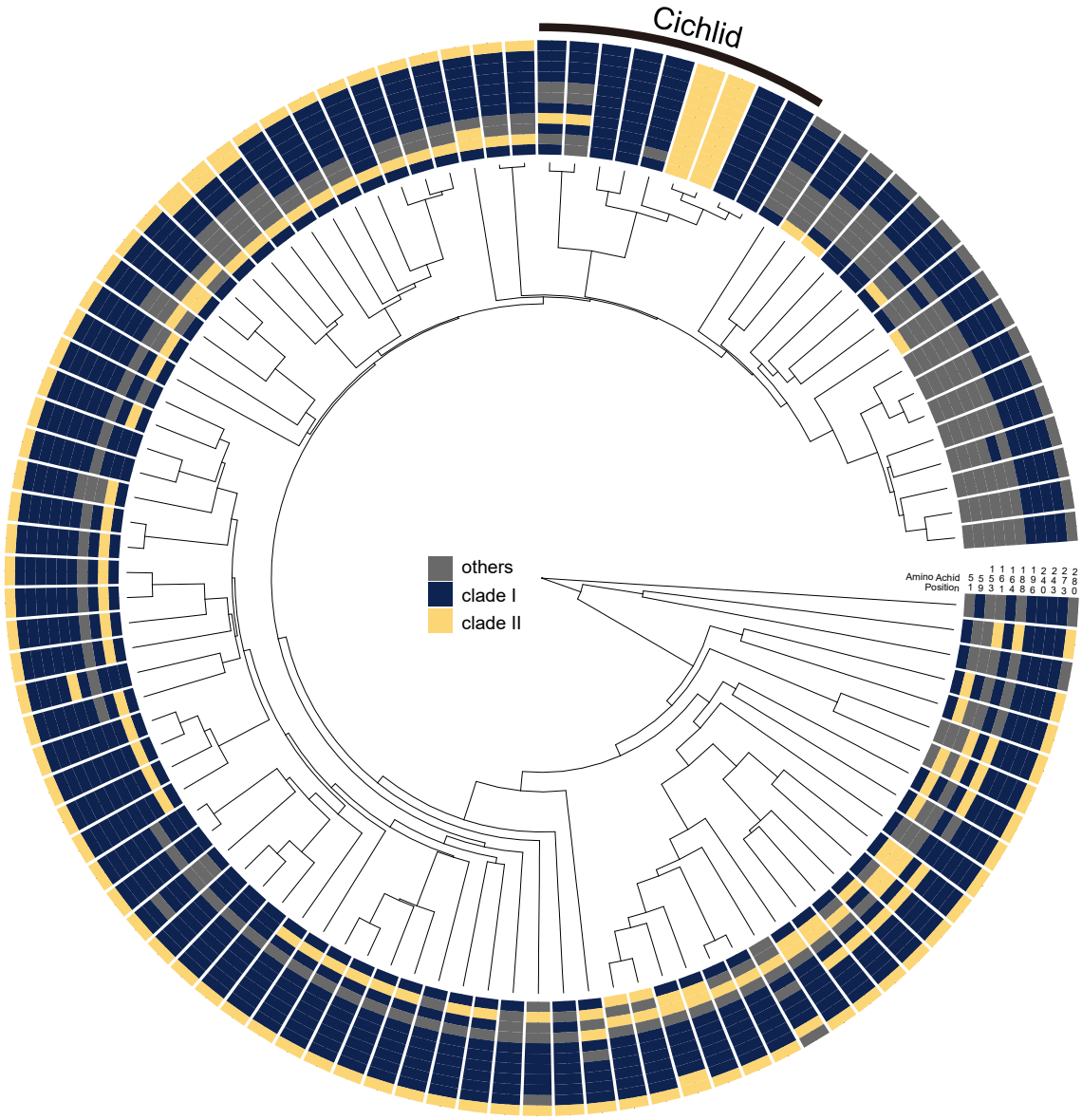


Figure S10 V1R1 alignment of 102 ray-finned fish species. The alignment displays sites corresponding to the two alleles (clade I and clade II). Sites matching clade I are highlighted in dark blue, those matching clade II are highlighted in yellow, and all other sites are shown in grey.

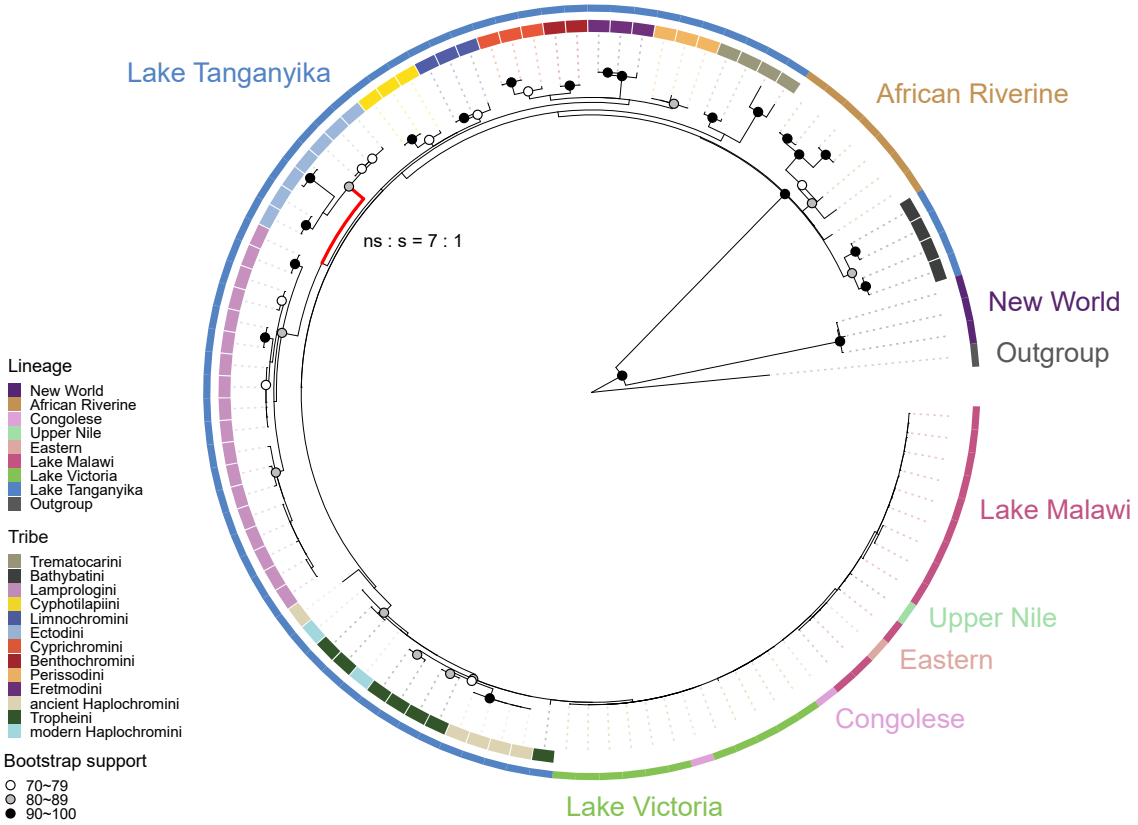


Figure S11 A phylogenetic tree of the V1R3 gene in cichlids. Nodes are marked with dots based on bootstrap values: black for values of 90 or higher, grey for 80–89, and white for 70–79. The inner circle represents tribes, while the outer circle indicates habitats. Branches under positive selection are highlighted in red (p=0.0361, see Table S2). The number of nonsynonymous and synonymous substitutions for the branch is indicated as ns:s.

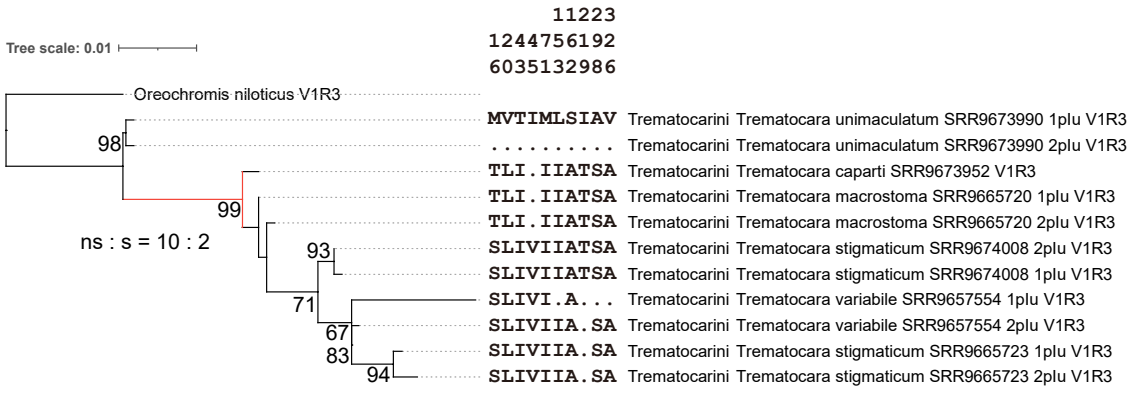


Figure S12 A phylogenetic tree and sequence alignment of the V1R3 gene in Trematocarini. Bootstrap values are shown only for nodes with support values of 60 or higher. *Trematocara unimaculatum*, which retains the ancestral sequence of V1R3, differ significantly from those of other species. The amino acid alignment displays only the sites with variations observed within Trematocarini. Dots indicate identity with the top sequence. Positive selection was not detected in the red branch (p=0.0503, see Table S2). The number of nonsynonymous and synonymous substitutions in the red branch is indicated as ns:s.

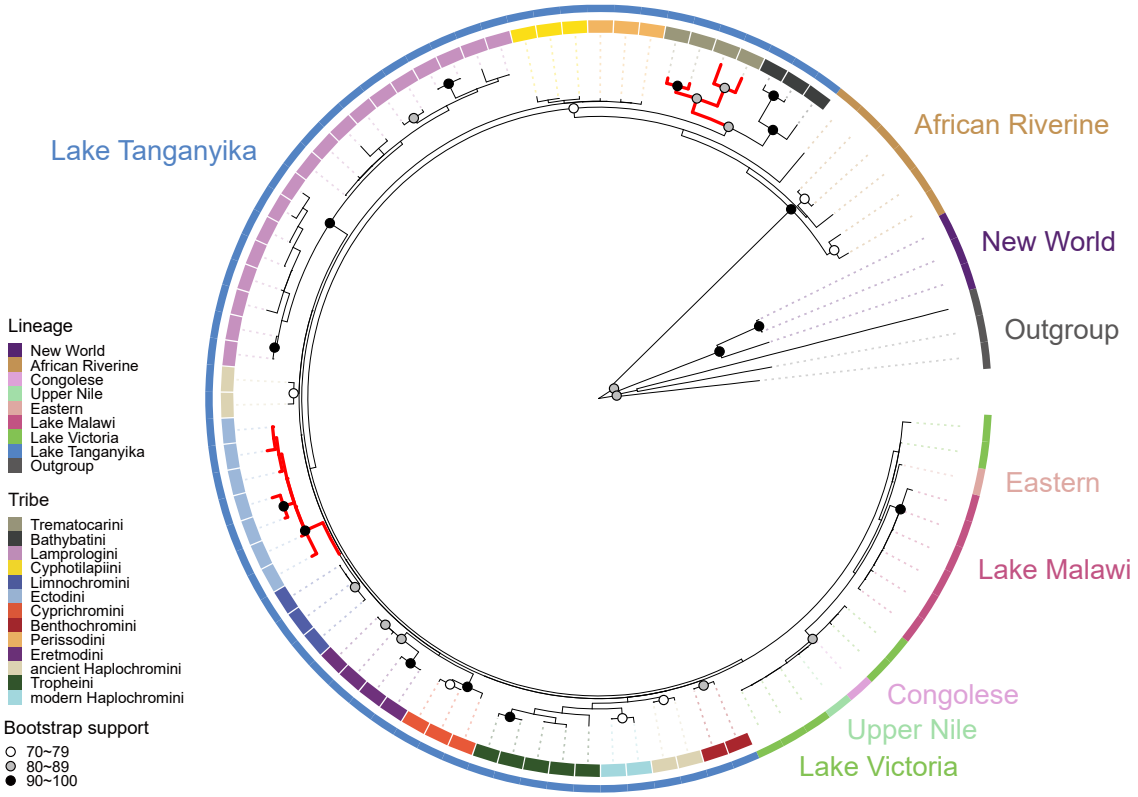


Figure S13 A phylogenetic tree of the V1R6 gene in cichlids. Nodes are marked with dots based on bootstrap values: black for values of 90 or higher, grey for 80–89, and white for 70–79. The inner circle represents tribes, while the outer circle indicates habitats. Branches under positive selection are highlighted in red (see Table S2).

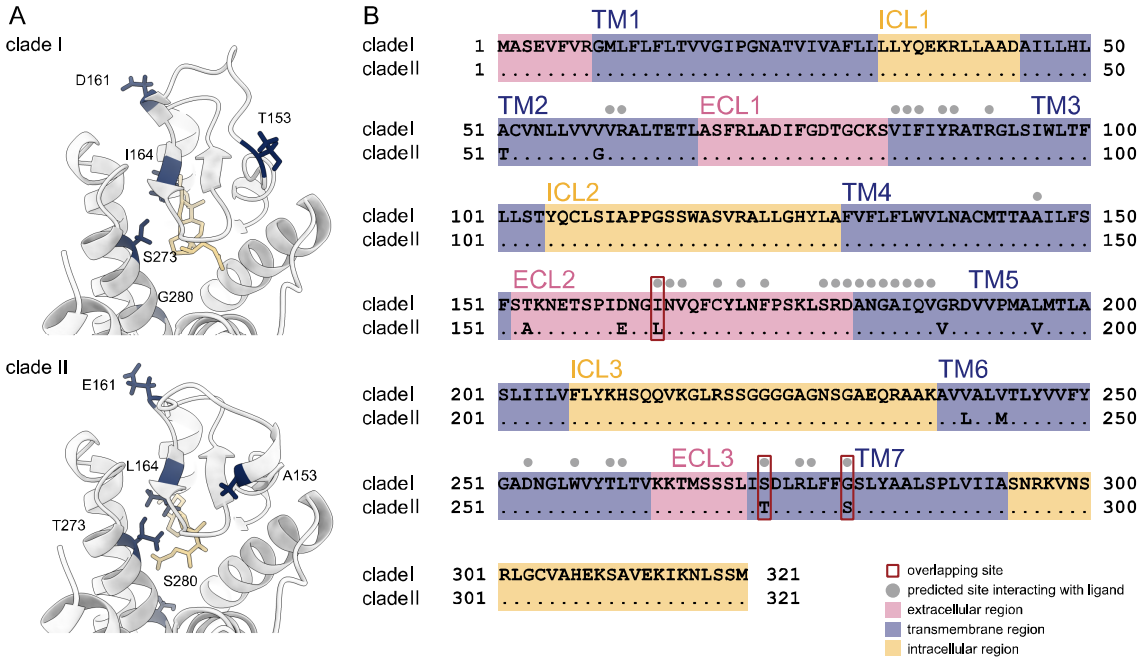


Figure S14 3D structure of cichlid V1R1. (A) 3D structure of V1R1 viewed from the extracellular side. Amino acids that differ between Clade I and Clade II are shown in dark blue. The ligand used for analysis (LCA) is shown in yellow. (B) Predicted transmembrane regions of cichlid V1R1. The transmembrane regions were predicted using the sequence of Clade I. Dots indicate amino acids identical to those in Clade I. The extracellular regions are shown in pink, transmembrane regions in blue, and intracellular regions in yellow. TM: transmembrane; ECL: extracellular loop; ICL: intracellular loop. Gray dots indicate predicted ligand-binding sites. Red squares indicate ligand-binding sites that are also sites of substitution between alleles. (Fisher's exact test: $p = 0.0922$)

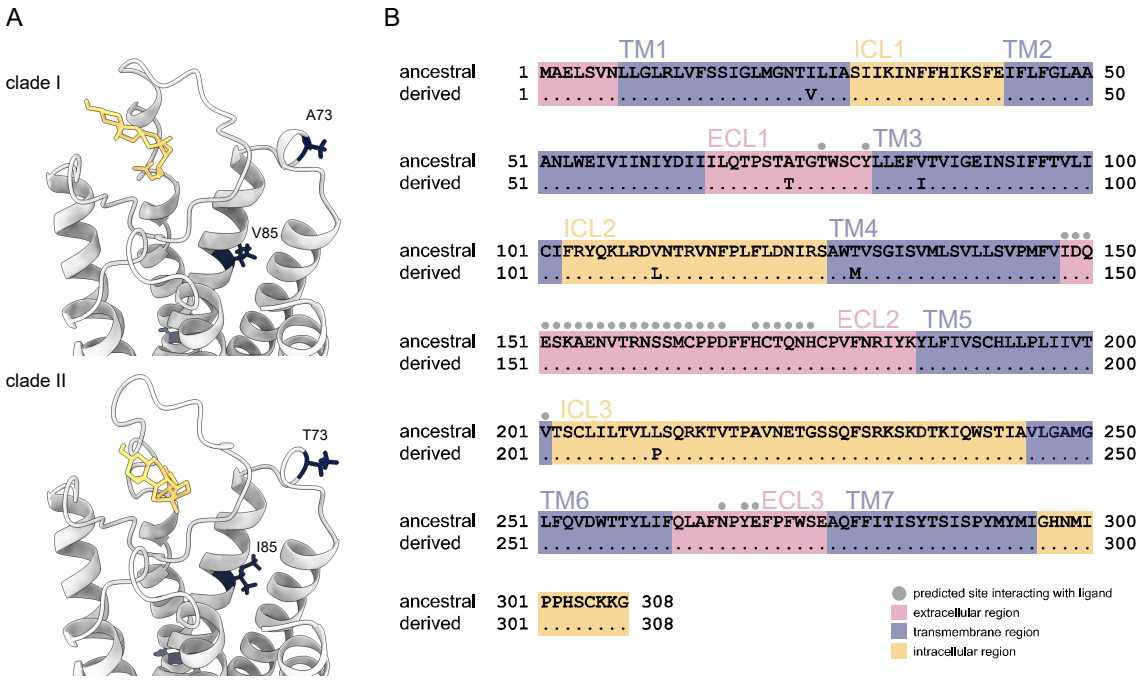
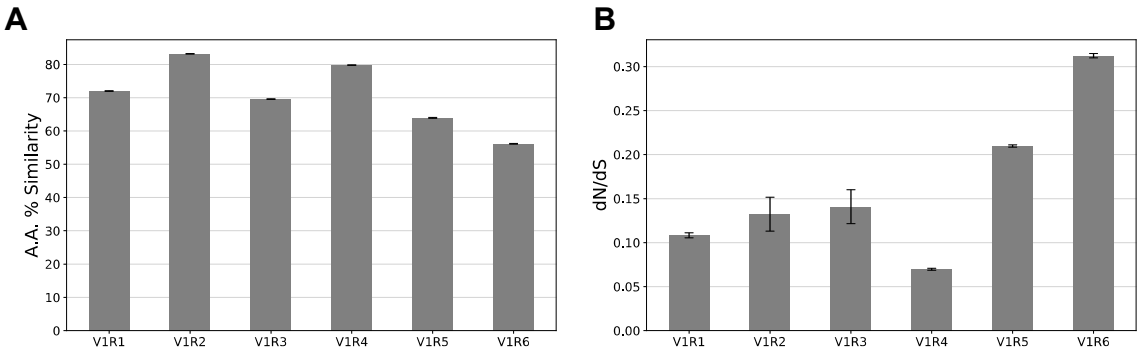


Figure S15 3D structure of cichlid V1R6. (A) 3D structure of V1R6 viewed from the extracellular side. Amino acids that differ between ancestral allele and derived allele are shown in dark blue. The ligand used for analysis (LCA) is shown in yellow. (B) Predicted transmembrane regions of cichlid V1R6. The transmembrane regions were predicted using the sequence of ancestral allele. Dots indicate amino acids identical to those in ancestral allele. The extracellular regions are shown in pink, transmembrane regions in blue, and intracellular regions in yellow. TM: transmembrane; ECL: extracellular loop; ICL: intracellular loop. Gray dots indicate predicted ligand-binding sites.

875 

876 Figure S16 Sequence similarity and dN/dS ratios of V1R genes among 102 ray-finned fish
877 species (A) Amino acid similarity of V1R genes. The average sequence similarity was
878 calculated for all pairwise combinations of the 102 species. Error bars represent
879 standard errors. (B) dN/dS ratios of V1R genes. The average dN/dS ratio was
880 calculated for all pairwise combinations of the 102 species. Error bars represent
881 standard errors.
882

883 Table S1 We used short-read data from 28 tribes, 528 species, and 908 samples of cichlids

884 registered in the NCBI Sequence Read Archive (SRA).

885

886 Table S2 The results of selection analyses and likelihood ratio tests.

V1R2

Model	np	lnL	p-value	omega
Null	66	-3334.240925		0.34224
Alternative	67	-3328.619231	0.000799	2.97063

V1R1

Model	np	lnL	p-value	omega
Null	61	-2086.808159		0.86632
Alternative	62	-2082.84435	0.004869	537.52213

V1R3

Model	np	lnL	p-value	omega
Null	208	-4602.837321		0.3905
Alternative	209	-4600.641708	0.036124	2.88326

V1R3 Trematocarini

Model	np	lnL	p-value	omega
Null	26	-1819.23301		0.50765
Alternative	27	-1817.317734	0.050326	2.13329

V1R6 Ectodini

Model	np	lnL	p-value	omega
Null	182	-5557.85997		0.54623
Alternative	183	-5553.243731	0.002378	2.6782

V1R6 Trematocarini

Model	np	lnL	p-value	omega
Null	182	-5557.85997		0.54591
Alternative	183	-5554.585293	0.010492	1.60894

887