Possible involvement of ghost introgressions in the striking diversity of Vomeronasal type 1 receptor genes in East African cichlids

Running title: Ghost Introgression V1R Diversity in Cichlid

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@life.isct.ac.jp

Abstract

Cichlids that have undergone adaptive radiation are genetically close but exhibit extreme ecological and morphological diversity, making them useful for understanding speciation mechanisms. Vomeronasal type 1 receptors (V1R) are highly conserved among teleost fish at the amino acid sequence level and believed to play a fundamental role in reproduction. We previously reported the surprisingly high sequence diversity of V1Rs among certain cichlid species, suggesting a possible role for V1Rs in their speciation. In this study, we investigated the process of evolutionary diversification of all 6 V1Rs (V1R1-6) by using the genome data of 528 cichlid species, encompassing nearly all lineages. In the case of V1R2, two highly divergent alleles (1.17%: variant sites/CDS length) without recombination were preserved and shared among cichlids found in all of the East African Great Lakes. In the case of V1R6, numerous highly variable alleles that could be derived from multiple recombination events between two highly divergent alleles (1.39%: variant sites/CDS length) were found among the Lake Victoria cichlids. Additionally, we identified highly divergent alleles of V1R1 within the tribe Tropheini, and of both V1R3 and V1R6 within Trematocarini and Ectodini. Because one of the two divergent alleles of these V1Rs emerged rapidly during cichlid evolution, they are likely to have been derived from introgression. However, despite extensive investigations, we could not identify the source lineages for these introgressions, implying that they may have become extinct. This study revealed the potential role of introgression in explaining the remarkable diversity of V1Rs in East African cichlids.

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

Introduction

Cichlids (family Cichlidae) are tropical freshwater fish that inhabit various regions spanning Central and South America to Africa, Madagascar, and India. The total number of cichlid species is estimated to be approximately 3,000, representing approximately 10% of all teleost fish (Kocher, 2004). Notably, approximately half of the extant cichlid species inhabit the East African Great Lakes, specifically Lake Tanganyika, Lake Malawi, and Lake Victoria (Salzburger et al., 2014), where they have emerged through large-scale adaptive radiation in each lake (Kocher, 2004; Salzburger, 2018). These cichlid species are widely used as model organisms for investigating speciation mechanisms, owing to their significant ecological and morphological diversity despite their genetic closeness. Cichlid species usually exhibit prezygotic isolation, ensuring their reproductive isolation from one another (Rometsch et al., 2020). Speciation is facilitated by sexual selection and assortative mating, particularly in environments where geographic isolation is less probable, such as lakes (Seehausen & Wagner, 2014).

Cichlids rely on various sensory systems for their reproductive behavior (Escobar-Camacho & Carleton, 2015). Because cichlids exhibit multiple body colors, their visual system is a subject of ongoing research. Previous studies have shown that the visual system has diversified according to ecological conditions and habitats, 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, termed “sensory drive,” plays a significant role in the speciation of cichlids. For instance, a shift in the peak sensitivity of long-wavelength-sensitive (LWS) opsin in response to the light environment (blue or red light), coupled with a divergence in female preference for male nuptial coloration (blue or

red), led to cichlid speciation in Lake Victoria (Seehausen et al., 2008).

Cichlids utilize visual and olfactory cues to recognize different species and individuals (Keller-Costa et al., 2015; Santos et al., 2023). Behavioral experiments on Lake Malawi cichlids have demonstrated the role of olfaction in conspecific recognition (Plenderleith et al., 2005). Additionally, olfactory imprinting during mouthbrooding was observed in Lake Victoria cichlids (Verzijden & ten Cate, 2007). Moreover, compounds that function as pheromones by influencing male behavior and the female reproductive system were found in the urine of Mozambique tilapia (*Oreochromis mossambicus*) (Ashouri et al., 2023; Keller-Costa et al., 2014). Therefore, cichlids largely depend on olfaction for conspecific recognition; however, in contrast to studies of the visual system, research on which olfactory receptor genes and alleles contribute to species recognition and, hence, assortative mating remains limited.

Four types of olfactory receptor genes—olfactory receptor, trace amine-associated receptor, vomeronasal type 1 receptor (V1R), and V2R—have been identified in teleost fish, all of which are expressed in the olfactory neurons (Calvo-Ochoa et al., 2021). Of these, the V1R gene, also called olfactory receptor related to class A (ORA), is considered to play a role in pheromone detection in fish, as it is homologous to the mammalian V1R, a known pheromone receptor gene (Del Punta et al., 2002; Pfister & Rodriguez, 2005). Although mammalian V1Rs exhibit copy number and sequence diversity, teleost fish V1Rs are highly conserved, with six copies of the gene (V1R1–6) shared across species, and their orthologs are also highly conserved at the amino acid sequence level (Saraiva & Korsching, 2007). Therefore, V1R likely detects hormonal pheromones shared across species, contributing to synchronizing gamete maturation and/or spawning interactions (Ota et al., 2012; Policarpo et al., 2022; Saraiva & Korsching, 2007; Zapilko & Korsching, 2016). The V1Rs of zebrafish

have been demonstrated to detect 4-hydroxyphenylacetic acid (4HPAA) and bile acids (Behrens et al., 2014; Cong et al., 2019). Additionally, V1R2- and V1R5-expressing neurons were activated in response to male cichlid urine (Kawamura & Nikaido, 2022). Therefore, V1Rs can significantly contribute to cichlid speciation; however, whether their genetic diversity is associated with speciation or environmental adaptation, as demonstrated in the case of LWS, is unclear.

The V1R genes exhibit sequence-level diversity in East African cichlids (Nikaido et al., 2014). We previously found that certain tribes of Lake Tanganyika cichlids exhibit allelic diversity in the V1R1, V1R3, and V1R6 genes. We also found two alleles of V1R2, and revealed that these alleles are present in cichlid species from different lakes. These alleles differed by >10 nonsynonymous substitutions, suggesting functional divergence. Hence, V1Rs are well-suited for investigating the possibility of olfaction-mediated speciation in cichlids (Santos et al., 2023). However, the limited number of species studied prevents a complete understanding of the origin and development of these divergent alleles.

The recent availability of whole-genome sequencing (WGS) data for >500 cichlid species has enabled detailed comparative genomic analyses across different species (Malinsky et al., 2018; McGee et al., 2020; Ronco et al., 2021). These data enabled us to extensively examine the V1R sequences to elucidate the origin and maintenance of divergent alleles across lineages in a range of morphologically and ecologically diverse cichlids. Accordingly, we conducted a comprehensive analysis of V1R allelic diversity using WGS data from 528 species and 907 specimens of cichlids. We also investigated how sequence-level variations affect the V1R function, particularly its ligand-binding function.

Materials and Methods

V1R gene sequences

We used short-read data from 28 tribes, 528 species, and 907 samples of cichlids deposited within the NCBI Sequence Read Archive (Figure S1, Table S1). Low-quality reads were removed using fastp v0.23.2 (Chen et al., 2018), and the filtered reads were mapped to the reference genome (Supplementary Information) using bwa-mem2 v2.2.1 (Vasimuddin et al., 2019) and sorted and indexed using samtools v.1.15 (Danecek et al., 2021). The V1R gene regions were then extracted from the sorted BAM files using samtools v1.15 (Danecek et al., 2021). Variants were called from the extracted BAM files using the HaplotypeCaller tool of GATK v4.3.0 (McKenna et al., 2010). The first round of hard filtering for SNPs and INDELs was performed using the VariantFiltration tool of GATK v4.3.0 (McKenna et al., 2010) based on the following parameters: SNPs (QD<2.0, QUAL<30.0, SOR>4.0, FS>60.0, MQ<40.0, MQRankSum<-12.5, and ReadPosRankSum<-8.0) and INDELs (QD<2.0, QUAL<30.0, FS>200.0, SOR>10.0, and ReadPosRankSum<-20.0). Based on the resulting VCF file, quality recalibration of the BAM files was performed with the BaseRecalibrator and ApplyBQSR tools from GATK v4.3.0 (McKenna et al., 2010). The second round of hard filtering was conducted similarly to the first round. The filtered VCF files were used to generate V1R gene sequences in fasta format using bcftools v1.15 (Danecek et al., 2021). Sequences containing heterozygous variants were phased using HapCUT2 v1.3.3 (Edge et al., 2017). We developed a custom pipeline spanning the process from short-read data download to obtaining V1R gene sequences.

Phylogenetic analysis

To minimize the computational load of the phylogenetic analysis, we removed duplicate

135 sequences and then randomly extracted 45 sequences from an initial pool of 337 V1R2
136 sequences, resulting in 5 datasets (Figure 1, Figure S2–6). The datasets were subjected to
137 multiple sequence alignment using MAFFT v7.520 (Katoh & Standley, 2013) with the
138 following options: —maxiterate 16 —localpair. Maximum likelihood trees were constructed
139 using RAxML-NG v1.2.0 (Kozlov et al., 2019) with 1000 bootstrap replicates. The HKY+G4
140 model was selected using Modeltest-NG v0.1.7 (Darriba et al., 2020) for phylogenetic tree
141 construction. Datasets 1 through 5 exhibited similar topological patterns (Figure S2–6); thus,
142 we only used Dataset 1 (Figure 1). For V1R1 of Tropheini, we used 43 nonduplicated V1R1
143 sequences and the outgroup *Oreochromis niloticus* V1R1 sequence, totaling 44 sequences
144 (1000 bootstrap replicates, model: HKY+I). For V1R3 of Trematocarini, we used 11
145 nonduplicated V1R3 sequences and the outgroup *Oreochromis niloticus* V1R1 sequence,
146 totaling 12 sequences (1000 bootstrap replicates, model: HKY+I). For both V1R3 and V1R6,
147 we used 53 nonduplicated sequences (100 bootstrap replicates, model V1R3: HKY+G4,
148 V1R6: GTR+G4). We used the following R packages for visualization: treeio (Wang et al.,
149 2020), ggtree (Yu et al., 2017), readr (Wickham H, Hester J, Bryan J, 2024), dplyr (Wickham
150 H, François R, Henry L, Müller K, Vaughan D, 2023), ggplot2 (Villanueva & Chen, 2019),
151 ggmsa (Zhou et al., 2022), and iTOL v6 (Letunic & Bork, 2024). We calculated the
152 frequencies of each V1R1 and V1R2 allele across cichlid lineages using unphased
153 sequences (V1R1: 907 sequences, V1R2: 907 sequences). We used 1309 sequences to
154 classify the V1R6 haplotypes. Additionally, we used V1R sequences from 102 species of
155 ray-finned fish (Policarpo et al., 2022) and phylogenetic tree data (Chang et al., 2019;
156 Rabosky et al., 2018) to determine the sequence diversity of V1R genes in ray-finned fish.
157 We generated a pairwise similarity matrix for the 102 species using CIAlign v1.0.17
158 (Tumescheit et al., 2022) and calculated the average similarity. We also calculated pairwise

159 dN/dS ratios using the CODEML program in PAML v4.10.6 (Yang, 2007) and computed the
160 average value.
161
162 **Selection analysis**
163 The amino acid sequences translated using Aliview (Larsson, 2014) were subjected to
164 multiple sequence alignment using MAFFT v7.520 (Katoh & Standley, 2013) with the
165 following options: —maxiterate 16 —localpair. The resultant sequence alignment was
166 converted to a codon alignment using PAL2NAL v14 (Suyama et al., 2006). We estimated
167 the dN/dS ratios for branches with an accumulation of nonsynonymous substitutions using
168 the branch model with the codeml program in PAML v4.10.7 (Yang, 2007). We then
169 performed a likelihood ratio test to compare the foreground and background branches and
170 assess the statistical significance of the dN/dS increase in the foreground branch. Clades I
171 and II were established within the V1R2 grouping. The dN/dS ratio was calculated across all
172 five datasets using PAML v4.10.7 (Datasets 1–5, Figure S2–6), and ancestral sequences
173 were reconstructed for the ancestral and descendant nodes of each branch. Analysis of the
174 reconstructed ancestral sequences revealed 11 nonsynonymous substitutions shared
175 across all datasets, which were classified as mutations distinguishing Clade I from Clade II.
176 For V1R1, we defined Clades I and II based on the phylogenetic tree of the V1R1 gene in
177 Tropheini (Figure S9). We identified 11 nonsynonymous substitutions, which were classified
178 as mutations distinguishing Clade I from Clade II.
179
180 **Protein modeling and ligand-binding predictions**
181 We modeled the 3D structures of the V1R1, V1R2, and V1R6 proteins using AlphaFold v2.3
182 (Jumper et al., 2021). We used the most reliable structures to estimate the coordinates of

183 the ligand-binding pockets using Deepsite (Jiménez et al., 2017) in PlayMolecule Viewer
184 (Torrens-Fontanals et al., 2024). We used the obtained coordinates to perform ligand
185 docking simulations in Webina v.1.0.5 (Kochnev et al., 2020) using 4HPAA, a known ligand
186 for zebrafish V1R2 (Behrens et al., 2014), and lithocholic acid, a known ligand for zebrafish
187 V1R1 and V1R6 (Cong et al., 2019). As a result, 5 relative positions (poses) of the ligand in
188 V1R1 (Clade I: 3 poses, Clade II: 2 poses), 18 poses in V1R2 (Clade I: 9 poses, Clade II: 9
189 poses), and 18 poses in V1R6 (ancestral allele: 9 poses, derived allele: 9 poses) were
190 predicted. Residues located within 5Å of all these poses were designated as ligand-related
191 sites. Furthermore, we predicted transmembrane regions using DeepTMHMM v1.0.39
192 (Hallgren et al., 2022). The results were visualized using ChimeraX v1.6 (Goddard et al.,
193 2018).
194
195 **Recombination rate**
196 We compared the relative recombination rates in the regions surrounding the V1R genes
197 using the recombination rate of *Pundamilia nyererei* (Feulner et al., 2018) and the genomic
198 position of the V1R genes. We obtained the genomic positions of the V1R genes using
199 BLASTN with the *Pundamilia nyererei* genome assembly (P_nyererei reference v2) (Feulner
200 et al., 2018) as the reference and *Oreochromis niloticus* V1R sequences from NCBI (V1R1,
201 AB671405.1; V1R2, AB670433.1; V1R3, AB704581.1; V1R4, AB704618.1; V1R5,
202 AB704661.1; and V1R6, AB704705.1) as the query. The genomic positions of the V1R1,
203 V1R2, V1R3, V1R4, V1R5, and V1R6 genes were identified as chr13:8171366-8172331,
204 chr13:8175817-8176758, chr13:18527399-18530109, chr13:18531988-18532072,
205 chr2:41585178-41586302, and chr7:27107196-27108122, respectively.

Results

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

We constructed a phylogenetic tree for the V1R2 gene using sequence data from 44 species, which were selected from all cichlid lineages (Figure 1). Cichlids in the East African Great Lakes have undergone independent adaptive radiation in each lake, resulting in closely related species in each lake, with species from Lake Malawi and Lake Victoria forming distinct monophyletic groups (Svardal et al., 2021). However, the topology of the V1R2 gene tree did not match the species tree and was divided into Clades I and II with a maximal bootstrap probability (Figure 1). Clade I included cichlids from the outgroup (India and Madagascar), the African Riverine, Lake Tanganyika, Lake Malawi, and Lake Victoria, whereas Clade II included cichlids from Lake Tanganyika, ancient Haplochromini, modern Haplochromini, Lake Malawi, Upper Nile lineage, Eastern lineage, Congolese lineage, and Lake Victoria. Notably both clades contained lineages from all three of the East African Great Lakes. Clades I and II were distinguished based on 11 amino acid substitutions (Figure 1, see Method). Furthermore, the dN/dS ratio of the branch at which Clade II diverged from Clade I was significantly higher than that of the background branches (Table S2), indicating a positive selection acting on the divergent branch.

We calculated the frequencies of the two major alleles (Clades I and II) of V1R2 for each cichlid lineage (Figure 2). In the Lamprologini and Ectodini, which inhabit Lake Tanganyika, V1R2 alleles were fixed in Clade II. In the Haplochromini from Lake Malawi and Lake Victoria, the alleles were not fixed in either clade, resulting in a mix of homozygotes of Clade I or II alleles, along with their corresponding heterozygotes. Despite extensive exploration, Clade II was not detected in the lineages from India, Madagascar, the New World, and the African Riverine. Notably, recombination was not found between the two

230 alleles in all lineages except Cyphotilapiini, despite the alleles coexisting as polymorphisms
231 within populations or species. The polymorphic sites in cichlids that distinguish Clade I from
232 Clade II were highly conserved among noncichlid ray-finned fish, with all V1R2 polymorphic
233 sites being classified into Clade I (Figure S7).
234
235 **Highly variable alleles of V1R6 with multiple recombination in Lake Victoria cichlids**
236 We constructed a V1R6 phylogenetic tree, and its topology was consistent with that of the
237 species tree (Figure 3A). We found the Lake Victoria cichlids to exhibit a notable number of
238 alleles, which were differentiated by a combination of 13 variant sites (Figure 3B). These
239 allelic variants were present as polymorphisms across different species. Substitutions
240 occurring independently multiple times at the same site for each allele are extremely unlikely.
241 Therefore, we hypothesized that this wide range of alleles originated from recombination
242 between two highly divergent alleles present in the ancestral population of Lake Victoria
243 (Figure 3C). Accordingly, we found that the recombination rate of the genomic region
244 surrounding V1R6, estimated using data from a previous study (Feulner et al., 2018), was
245 higher than that of regions surrounding other V1R genes (Figure S8). Based on our
246 hypothesis, we focused on the 13 variant sites identified in Lake Victoria cichlids and
247 subsequently classified the V1R6 alleles of cichlids present across all the East African Great
248 Lakes and rivers. We did not find any potential signature of recombination between highly
249 divergent alleles in cichlids residing outside the Lake Victoria basin. Subsequently, we
250 reconstructed the ancestral sequence at the node denoted by the arrowhead in Figure 3A
251 and designated this sequence as the “ancestral allele” (Figure 3B, left top; dark blue
252 nucleotide). We then subtracted the nucleotides of the “ancestral allele” (dark blue) from the
253 13 polymorphic sites of Lake Victoria cichlids to reconstruct a hypothetical “derived allele”

254 (yellow), representing a counterpart sequence of the past recombination event (Figure 3B,
255 right top). Accordingly, the alleles found in Lake Victoria cichlids were categorized as
256 “ancestral,” “derived,” and their “mixtures.” Notably, cichlids from the Upper Nile, Eastern
257 lineage, and Congolese lineage possess recombinant alleles similar to those of Lake
258 Victoria cichlids (Figure 3B). However, despite extensive investigations, the hypothetical
259 “derived allele” was not found in any extant cichlid, except for those in Lake Victoria (e.g.,
260 haplotypes Lake Victoria 51 and 52; Figure 3B, right bottom).
261
262 **Highly divergent alleles of V1R1, V1R3, and V1R6 in cichlids of particular tribes**
263 We identified two highly divergent Tropheini V1R1 alleles, namely Clades I and II, which are
264 distinguished based on 11 amino acids (Figure 4A, see Method). The monophyly of Clades I
265 and II was corroborated by a maximal bootstrap probability (Figure S9). The amino acid
266 sequence of Tropheini Clade I is similar to that of other cichlids; thus, Clades I and II are
267 defined as “ancestral” and “derived” alleles, respectively. Furthermore, the dN/dS ratio of the
268 branch where Clade II diverged from Clade I was significantly higher than that of the
269 background branches (Table S2), indicating a positive selection acting on the divergent
270 branch. We then calculated the frequencies of the two major V1R1alleles (Clades I and II) in
271 Tropheini and found that Clades I and II were present in 28% and 65% of the Tropheini
272 samples, respectively (Figure 4B). Extensive investigations of Clade II sequences revealed
273 their presence only in Tropheini cichlids among the various cichlids examined. The
274 polymorphic sites differentiating Clade I from Clade II in cichlids were highly conserved in
275 Clade I of ray-finned fish (Figure S10).
276 We also identified highly divergent alleles of V1R3 in the Ectodini and Trematocarini.
277 Seven nonsynonymous substitutions were identified on the branch corresponding to the

278 common ancestor of Ectodini (Figure S11). This branch exhibited a significantly higher
279 dN/dS ratio than the background branches (Table S2). Furthermore, 10 nonsynonymous
280 substitutions were found on the branch representing the common ancestor of Trematocarini,
281 excluding *Trematocara unimaculatum* (Figure S12). This branch exhibited a significantly
282 higher dN/dS ratio than the background branches (Table S2). We found divergent alleles for
283 V1R6 in the Ectodini and Trematocarini (Figure S13), with the dN/dS ratio of these clades
284 (Ectodini, omega = 2.69; Trematocarini, omega = 1.61) being significantly higher than that of
285 the background (Table S2).
286
287 **V1R2 amino acid substitution sites are located near the ligand-binding pocket**
288 To understand the functional impact of the amino acid substitutions observed between the
289 two major divergent alleles of V1R2, we modeled the 3D structure of the V1R2 protein and
290 predicted its ligand-binding sites (Figure 5A). We first modeled the 3D structures using the
291 amino acid sequences of Clades I and II and then used 4HPAA, a highly promising
292 candidate ligand (Behrens et al., 2014; Kawamura & Nikaido, 2022), to determine the
293 ligand-binding sites for each clade and compare them. Eight of the 11 amino acid
294 substitution sites (TM2: C62→Ser, ECL1: Q65→Pro, TM4: A145→Ser and A146→Thr,
295 ECL2: P159→Ser and H169→Leu, TM5: V185→Ile, and TM6: D247→Glu) we identified
296 were situated near the predicted ligand pocket (Figure 5B). Furthermore, four of these sites
297 (TM4: A145→Ser and A146→Thr, TM5: V185→Ile, and TM6: D247→Glu) overlapped with
298 the ligand-binding sites (Figure 5B). Additionally, the amino acid substitutions of V1R2 were
299 significantly skewed toward the ligand-binding sites (Fisher's exact test: p = 0.0389).
300 We also performed the same analyses for V1R1 and V1R6 (Figure S14A, S15A). In V1R1,
301 5 of the 11 amino acid substitution sites (TM2: V59→Gly, ECL2: I164→Leu, TM5: G188→

302 Val, and TM7: S273→Thr and G280→Ser) we identified were situated close to the predicted
303 ligand pocket (Figure S14B). Furthermore, three of these sites (ECL2: I164→Leu; TM7:
304 S273→Thr and G280→Ser) overlapped with the ligand-binding sites (Figure S14B).
305 However, the skewed distribution of the V1R1 amino acid substitutions toward the
306 ligand-binding sites did not show statistical significance (Fisher's exact test: $p = 0.0922$). In
307 V1R6, none of the amino acid substitution sites were located close to the predicted ligand
308 pocket (Figure S15B).

309 **Discussion**

310 **Allelic and functional diversity of V1R in East African cichlids**

311 We conducted large-scale analyses of V1R diversity in cichlids using WGS data from 528
312 species and 907 samples, which have recently been made available following large-scale
313 comparative genomic studies (Malinsky et al., 2018; McGee et al., 2020; Ronco et al., 2021).
314 As a result, we identified an accumulation of nonsynonymous substitutions and signatures
315 of positive selection in V1R1, V1R2, V1R3, and V1R6 (Figure 6A). Notably, two highly
316 divergent alleles of V1R1 and V1R2 were found in the Tropheini (Figure 4B) and those
317 inhabiting the East African Great Lakes (Figure 2B), respectively. In addition, highly
318 divergent alleles of V1R3 and V1R6 were identified in the Trematocarini and Ectodini
319 (Figure S11–13). These results corroborate previous research involving small-scale species
320 sampling efforts (Nikaido et al., 2014).

321 The V1R gene orthologs are highly conserved across different fish species (Ota et al.,
322 2012; Policarpo et al., 2022; Saraiva & Korsching, 2007; Zapilko & Korsching, 2016). The
323 recently published V1R gene sequences of ray-finned fish (Policarpo et al., 2022) also
324 exhibit high sequence similarity and low dN/dS ratios (Figure S16). Furthermore, the
325 polymorphic sites that differentiate the divergent alleles of V1R1 and V1R2 in cichlids are
326 highly conserved among ray-finned fish (Figure S7, S10). These results suggest that V1R
327 gene diversification is unique to cichlids within the ray-finned fish category and is probably
328 associated with rapid adaptive radiation.

329 The prediction of the 3D structure of V1R2 and the determination of the sites interacting
330 with the ligand revealed that the amino acid residues differing between Clades I and II are
331 located close to the ligand pocket. Furthermore, amino acid substitutions were
332 predominantly found near the ligand-binding sites, showing a statistically significant trend

333 (Figure 5B). Two amino acid substitutions were detected in ECL2 (extracellular loop 2)
334 (Figure 5B), which is critical for ligand selectivity in class A G protein-coupled receptors
335 (GPCRs), a category that encompasses V1R (Nicoli et al., 2022). ECL2 also contributes to
336 ligand binding in the bitter taste receptor T2R, which is phylogenetically close to V1R (Xu et
337 al., 2022). These findings suggest that the two divergent clades of V1R2 likely differ from
338 one another in terms of their ligand-binding properties.

339 V1R6 exhibited the lowest sequence similarity and highest dN/dS ratio among the six
340 V1R genes in all ray-finned fish species (Figure S16). Hence, the pressure of negative
341 selection operating on V1R6 is weaker than that on other V1Rs, not only in cichlids but also
342 across all ray-finned fish species. Notably, V1R6 of the Lake Victoria basin cichlids exhibited
343 high variability, with numerous alleles resulting from multiple recombination events between
344 divergent alleles (Figure 3B, 3C). The recombination rate in the V1R6 region of Lake
345 Victoria cichlids was considerably high (Figure S8), indicating that V1R6 may be located in a
346 recombination hotspot. In contrast, recombination in V1R1 and V1R2 is strongly suppressed.
347 The structural constraint on V1R6 appears relatively weak, implying that the ligand
348 specificity of V1R6 may be lower than that of other V1Rs. In the 3D structural analysis, the
349 ligand pockets of V1R1 and V1R2 were predicted to be within the TM region, which is
350 analogous to that found in typical class A GPCRs (Figure 5B, S14B). Conversely, the ECL2
351 region was predicted to facilitate ligand binding in V1R6 (Figure S15B). This suggests that
352 V1R6 is distinct from V1R1 and V1R2 in terms of its functions and ligand binding
353 mechanisms. Investigating the distinct structural differences predicted among V1Rs will
354 provide insight into the function of V1Rs in fish, a topic of ongoing debate (Behrens et al.,
355 2014; Cong et al., 2019; Saraiva & Korsching, 2007).
356

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

358 Understanding the remarkable diversity of V1Rs in cichlids is crucial, particularly their origin
359 and maintenance over the 10 million years of their evolutionary history. We first investigated
360 the origin and process of V1R2 diversification by extensively investigating the existence of
361 Clade I and II alleles in each cichlid lineage (Figure 2). Clade II was found in cichlid lineages
362 inhabiting the East African Great Lakes but not in lineages that diverged prior to the Lake
363 Tanganyika radiation. Specifically, the Lamprologini and Ectodini were fixed into Clade II,
364 whereas those of other Lake Tanganyika tribes were fixed into Clade I. The cichlids of
365 Haplochromini were homozygotes of Clade I or II, or heterozygotes of both. The pattern of
366 V1R2 allele distribution indicates that Clade II emerged suddenly approximately 10 million
367 years ago, coinciding with the radiation of Lake Tanganyika cichlids. Given that the adaptive
368 radiation of cichlids in Lake Tanganyika was rapid, it is unlikely for 11 nonsynonymous
369 substitutions to accumulate gradually in a single allele within this timeframe, even under the
370 influence of positive selection. Therefore, it is parsimonious to assume that the Clade II
371 allele was already present in some populations/species before the radiation event and was
372 subsequently introgressed into Lake Tanganyika cichlids through hybridization.

373 The genomes of East African cichlids contain several examples of standing genetic
374 variation (SGV) (Nakamura et al., 2021; Svardal et al., 2021), including some that date back
375 to before the adaptive radiation events in Lake Tanganyika (Nakamura et al., 2023). Clade I
376 and II alleles of V1R2 represent clear examples of this old SGV. One of the most plausible
377 explanations for the origin of this SGV is the melting-pot hypothesis (Nakamura et al., 2023;
378 Weiss et al., 2015). According to this hypothesis, several cichlid populations/species,
379 previously scattered throughout wetlands and small lakes, hybridized just before the
380 adaptive radiation event in Lake Tanganyika. If each scattered population became fixed for

381 Clades I and II, their hybridization may have provided the origin of the SGV for the highly
382 divergent V1R2 alleles. However, one of the two alleles comprising the SGV of V1R2, Clade
383 II, was not detected despite an extensive investigation of the 46 basal lineages of the
384 riverine species (Figure 2). The absence of the Clade II allele in the riverine species
385 suggests that the transporter lineage of the Clade II allele is already extinct.

386 Gene flow from extinct lineages, called ghost introgression, has been identified in
387 various vertebrates, from gobiid fish to elephants, in recent years (Ai et al., 2015; Frei et al.,
388 2022; Gopalakrishnan et al., 2018; Kato et al., 2023; Palkopoulou et al., 2018; Zhang et al.,
389 2019). The sudden emergence of Clade II of V1R2 in Lake Tanganyika cichlids can be
390 attributed to ghost introgression. Introgressed alleles are usually identified as highly
391 divergent from those in nonintrogressed genomic regions. However, neutral genomic
392 regions frequently undergo recombination and genetic drift following introgression, thereby
393 complicating the identification of introgression signatures. Accordingly, V1R2 serves as a
394 good example for identifying introgression due to its strict constraints on recombination.
395 African Riverine lineages that predate adaptive radiation are referred to as "transporters" of
396 genetic diversity (Loh et al., 2013). Notably, gene flow from Steatocranini, a riverine lineage
397 that diverged before adaptive radiation (Irisarri et al., 2018), coincided with the emergence
398 of Clade II of V1R2. Therefore, the high diversity of V1R2 alleles (Clade II) is likely a result of
399 an introgression event from an extinct African Riverine lineage that diverged before adaptive
400 radiation (ghost introgression) or from unexamined lineages (Figure 6B).

401 Our findings based on V1R6 data suggest that a significant number of distinct alleles
402 emerged from the repeated recombination of 2 highly divergent alleles, one ancestral and
403 the other derived, in the cichlids of the Lake Victoria basin, which were differed by 13
404 nucleotide substitutions (Figure 3B). Recombination of the LWS opsin gene was observed in

405 Lake Victoria cichlids, which was attributed to hybridization between the Congo and Upper
406 Nile lineages (Meier et al., 2017). In contrast, since both the Congo and Upper Nile lineages
407 exhibit recombinant sequences, recombination between the ancestral and derived alleles of
408 V1R6 must have occurred earlier than the hybridization event of the Congo and Upper Nile
409 lineages. However, all the extant Haplochromini species, except Lake Victoria cichlids, lack
410 the derived allele. Since the estimated mutation rate in cichlids is 3.5×10^{-9} (per site per
411 generation), it is unrealistic for 13 substitutions to occur within the short timeframe of the
412 emergence of the cichlids of Lake Victoria basin (~400,000 years ago) (Johnson et al., 1996;
413 Salzburger et al., 2014). Therefore, the derived allele of V1R6 was likely to have been
414 introgressed from an extinct lineage (Figure 6B). The highly divergent alleles of V1R1 in
415 Tropheini, as well as those of V1R3 and V1R6 in Ectodini and Trematocarini, are probably
416 due to ghost introgression, similar to the cases of V1R2 and V1R6.

417 Notably, the nucleotide sequence differences between the two divergent alleles of
418 V1R1, V1R2, and V1R6 were mostly >1% (V1R1 = 1.14%, V1R2 = 1.17%, and V1R6 =
419 1.39%). Therefore, each divergent allele may have been introduced into the ancestral
420 population through a single ghost introgression event (melting-pot), following which the
421 derived alleles would have independently sorted into separate lineages. Alternatively,
422 introgression events in each lineage may have introduced the derived alleles of V1R1, V1R2,
423 and V1R6 (Figure 6A, 6B).

Conclusion

This study involved a comprehensive analysis of the V1R genes in East African cichlid species spanning numerous lineages, and we identified large-scale SGVs that can be attributed to ghost introgression. However, further research involving genome-wide statistical analyses rather than gene-level investigations of ghost introgression and melting-pot scenarios will be required to validate our findings. Our findings also highlighted the differences in structural constraints and selection pressure among the different V1Rs in cichlids, suggesting a functional distinction. This study provides important insights into the potential for V1R-mediated speciation and the genetic basis for adaptive evolution in cichlids.

Acknowledgments

This study was supported by JSPS KAKENHI (20KK0167, 24K02074, 24KK0141) to M.N. Phylogenetic analyses were partially performed using the NIG supercomputer at ROIS National Institute of Genetics.

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.

Conflict of Interest Statement

The authors declare no conflict of interest.

448 **Data Availability Statement**

449 V1R gene sequences of cichlids are available on Dryad (DOI). Scripts for obtaining V1R

450 gene sequences of cichlids are available on Github (https://github.com/taki-sh/RegionCall).

451

452 **Reference**

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

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

461 Behrens, M., Frank, O., Rawel, H., Ahuja, G., Potting, C., Hofmann, T., Meyerhof, W., &
462 Korsching, S. (2014). ORA1, a zebrafish olfactory receptor ancestral to all
463 mammalian V1R genes, recognizes 4-hydroxyphenylacetic acid, a putative
464 reproductive pheromone. *The Journal of Biological Chemistry*, *289*(28), 19778–
465 19788. https://doi.org/10.1074/jbc.M114.573162

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

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

473 Chang, J., Rabosky, D. L., Smith, S. A., & Alfaro, M. E. (2019). An r package and online
474 resource for macroevolutionary studies using the ray-finned fish tree of life. *Methods*
475 *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

572 Loh, Y.-H. E., Bezault, E., Muenzel, F. M., Roberts, R. B., Swofford, R., Barluenga, M., Kidd,
573 C. E., Howe, A. E., Di Palma, F., Lindblad-Toh, K., Hey, J., Seehausen, O.,
574 Salzburger, W., Kocher, T. D., & Streelman, J. T. (2013). Origins of shared genetic
575 variation in African cichlids. *Molecular Biology and Evolution*, *30*(4), 906–917.
576 https://doi.org/10.1093/molbev/mss326
577 Malinsky, M., Svardal, H., Tyers, A. M., Miska, E. A., Genner, M. J., Turner, G. F., & Durbin,
578 R. (2018). Whole-genome sequences of Malawi cichlids reveal multiple radiations
579 interconnected by gene flow. *Nature Ecology & Evolution*, *2*(12), 1940–1955.
580 https://doi.org/10.1038/s41559-018-0717-x
581 Matschiner, M., Böhne, A., Ronco, F., & Salzburger, W. (2020). The genomic timeline of
582 cichlid fish diversification across continents. *Nature Communications*, *11*(1), 5895.
583 https://doi.org/10.1038/s41467-020-17827-9
584 McGee, M. D., Borstein, S. R., Meier, J. I., Marques, D. A., Mwaiko, S., Taabu, A., Kishe, M.
585 A., O'Meara, B., Bruggmann, R., Excoffier, L., & Seehausen, O. (2020). The
586 ecological and genomic basis of explosive adaptive radiation. *Nature*, *586*(7827),
587 75–79. https://doi.org/10.1038/s41586-020-2652-7
588 McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella,
589 K., Altshuler, D., Gabriel, S., Daly, M., & DePristo, M. A. (2010). The Genome
590 Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA
591 sequencing data. *Genome Research*, *20*(9), 1297–1303.
592 https://doi.org/10.1101/gr.107524.110
593 Meier, J. I., Marques, D. A., Mwaiko, S., Wagner, C. E., Excoffier, L., & Seehausen, O.
594 (2017). Ancient hybridization fuels rapid cichlid fish adaptive radiations. *Nature*
595 *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

Tables and Figures

Figure 1 Phylogenetic tree and sequence alignment of the V1R2 gene in cichlids. Bootstrap values are shown only for nodes with support values of ≥60. The sequences are divided into two highly divergent alleles (Clades I and II). Only diagnostic amino acid sites between Clades I and II are 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, are excluded. Note that sequences from each lake (Tanganyika, Malawi, and Victoria) do not form monophyletic groups.

Figure 2 Frequencies of each allele of V1R2 in each cichlid lineage. We calculated the frequencies of Clade I and II alleles in each lineage. Clade I alleles are shown in dark blue, Clade II in yellow, heterozygotes in pink, and recombinants in gray. 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 represent the number of samples. The star represents the estimated acquisition times for 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. (B) All haplotypes of the V1R6 gene. Only sites that are polymorphic in Lake Victoria lineage are displayed. Numbers in parentheses following lineage names indicate the number of samples. Asterisks above the site numbers denote nonsynonymous substitutions. Ancestral and derived alleles at each site are 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 the recombination.

Figure 4 Two highly divergent alleles of V1R1 observed in Tropheini. (A) A phylogenetic tree of the V1R1 gene. Bootstrap values are shown only for nodes with support values of ≥60. (B) Frequencies of each allele of the V1R1 gene for each cichlid lineage. Clade I alleles are shown in dark blue, Clade II alleles in yellow, heterozygotes in pink, and recombinants in gray. 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 represent the number of samples. The star represents the estimated acquistion timing for Clade II.

Figure 5 3D structure of cichlid V1R2. (A) 3D structure of V1R2 viewed from the extracellular side. Amino acids that differ between Clades I and 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 Clade I sequence. 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 the predicted ligand-binding sites. Red squares indicate ligand-binding sites that are also sites of substitution between alleles. Mutations occurred at ligand-binding sites at a significantly higher rate than that expected by chance (Fisher's exact test: $p = 0.0389$).

773

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

783

Possible involvement of ghost introgressions in the striking diversity of Vomeronasal type 1 receptor genes in East African cichlids

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

Tables and Figures

Figure S1 Lineage names in this study are 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 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. The left panel shows nonsynonymous and synonymous substitution sites at positions marked with an asterisk.

Figure S3 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. The left panel shows nonsynonymous and synonymous substitution sites at positions marked with an asterisk.

Figure S4 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. The left panel shows nonsynonymous and synonymous substitution sites at positions marked with an asterisk.

Figure S5 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. The left panel shows nonsynonymous and synonymous substitution

sites at positions marked with an asterisk.

Figure S6 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. The left panel shows nonsynonymous and synonymous substitution sites at positions marked with an asterisk.

Figure S7 V1R2 alignment for 102 ray-finned fish species. The alignment displays sites corresponding to the two alleles (Clades I and 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 gray.

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 Phylogenetic tree of the V1R1 gene in Tropheini. Bootstrap values are shown only for nodes with support values of ≥60. The sequences are divided into two highly divergent alleles (Clades I and II). Recombination-derived sequences have been excluded.

Figure S10 V1R1 alignment for 102 ray-finned fish species. The alignment displays sites corresponding to the two alleles (Clades I and 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 gray.

Figure S11 Phylogenetic tree of the V1R3 gene in cichlids. Nodes are marked with dots based on bootstrap values: black for values of ≥90, gray for 80–89, and white for 70–79. The inner circle represents tribes, whereas 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 Phylogenetic tree and sequence alignment of the V1R3 gene in Trematocarini. Bootstrap values are shown only for nodes with support values of ≥60. *Trematocara unimaculatum*, which retains the ancestral sequence of V1R3, differs 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 Phylogenetic tree of the V1R6 gene in cichlids. Nodes are marked with dots based on bootstrap values: black for values of ≥90, gray for 80–89, and white for 70–79. The inner circle represents tribes, whereas 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 Clades I and II are shown in dark blue. The ligand used for analysis (lithocholic acid) is shown in yellow. (B) Predicted transmembrane regions of cichlid V1R1. The transmembrane regions were predicted using the Clade I sequence. 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 the 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 and derived alleles are shown in dark blue. The ligand used for analysis (lithocholic acid) is shown in yellow. (B) Predicted transmembrane regions of cichlid V1R6. The transmembrane regions were predicted using the ancestral allele sequence. Dots indicate amino acids identical to those in the 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 the predicted ligand-binding sites.

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

ratio was calculated for all pairwise combinations among the 102 species. Error bars represent the standard errors.

Table S1 Short-read data from 28 tribes, 528 species, and 908 samples of cichlids registered in the NCBI Sequence Read Archive (SRA).

Table S2 Selection analyses and likelihood ratio test results.

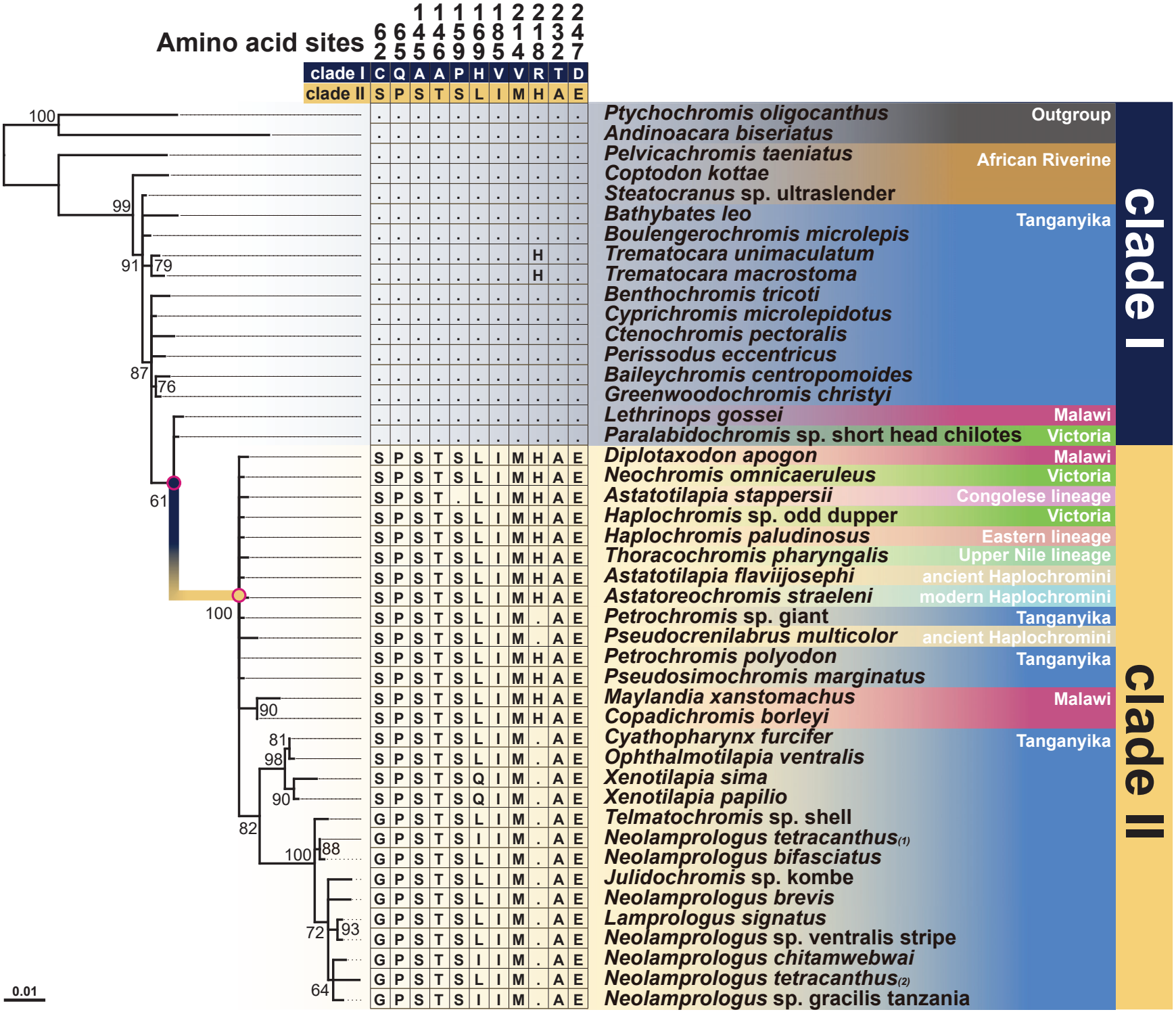

Amino acid sites
62 65 145 146 159 169 185 214 218 232 247
clade I C Q A A P H V V R T D
clade II S P S T S L I M H A E
Ptychochromis oligocanthus
Andinoacara biseriatus
Pelvicachromis taeniatus
Coptodon kottae
Steatocranus sp. ultraslender
Bathybates leo
Boulengerochromis microlepis
Trematocara unimaculatum
Trematocara macrostoma
Benthochromis tricoti
Cyprichromis microlepidotus
Ctenochromis pectoralis
Perissodus eccentricus
Baileychromis centropomoides
Greenwoodochromis christyi
Lethrinops gossei
Paralabidochromis sp. short head chilotes
Diplotaxodon apogon
Neochromis omnicaeruleus
Astatotilapia stappersii
Haplochromis sp. odd dupper
Haplochromis paludinosus
Thoracochromis pharyngalis
Astatotilapia flaviijosephi
Astatoreochromis straeleni
Petrochromis sp. giant
Pseudocrenilabrus multicolor
Petrochromis polyodon
Pseudosimochromis marginatus
Maylandia xanstomachus
Copadichromis borleyi
Cyathopharynx furcifer
Ophthalmotilapia ventralis
Xenotilapia sima
Xenotilapia papilio
Telmatochromis sp. shell
Neolamprologus tetracanthus(1)
Neolamprologus bifasciatus
Julidochromis sp. kombe
Neolamprologus brevis
Lamprologus signatus
Neolamprologus sp. ventralis stripe
Neolamprologus chitamwebwai
Neolamprologus tetracanthus(2)
Neolamprologus sp. gracilis tanzania
Outgroup
African Riverine
Tanganyika
Malawi
Victoria
Congolese lineage
Eastern lineage
Upper Nile lineage
ancient Haplochromini
modern Haplochromini
clade I
clade II
100
99
91
79
87
76
61
100
90
81
98
90
82
100
88
72
93
64
0.01

Figure 1

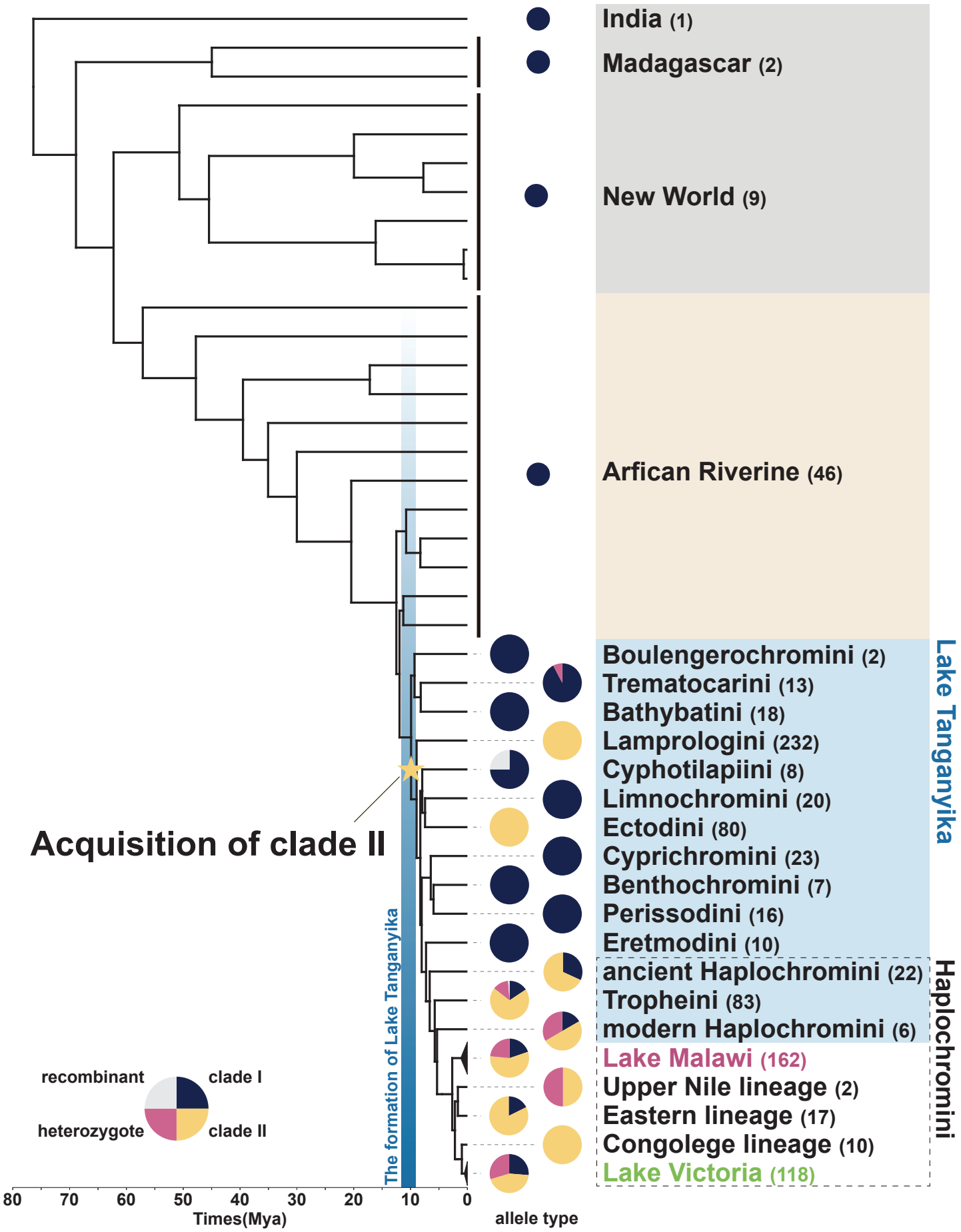

India (1)
Madagascar (2)
New World (9)
Arfican Riverine (46)
Boulengerochromini (2)
Trematocarini (13)
Bathybatini (18)
Lamprologini (232)
Cyphotilapiini (8)
Limnochromini (20)
Ectodini (80)
Cyprichromini (23)
Benthochromini (7)
Perissodini (16)
Eretmodini (10)
ancient Haplochromini (22)
Tropheini (83)
modern Haplochromini (6)
Lake Malawi (162)
Upper Nile lineage (2)
Eastern lineage (17)
Congolege lineage (10)
Lake Victoria (118)
Lake Tanganyika
Haplochromini
Acquisition of clade II
The formation of Lake Tanganyika
recombinant
clade I
heterozygote
clade II
80
70
60
50
40
30
20
10
0
Times(Mya)
allele type

Figure 2

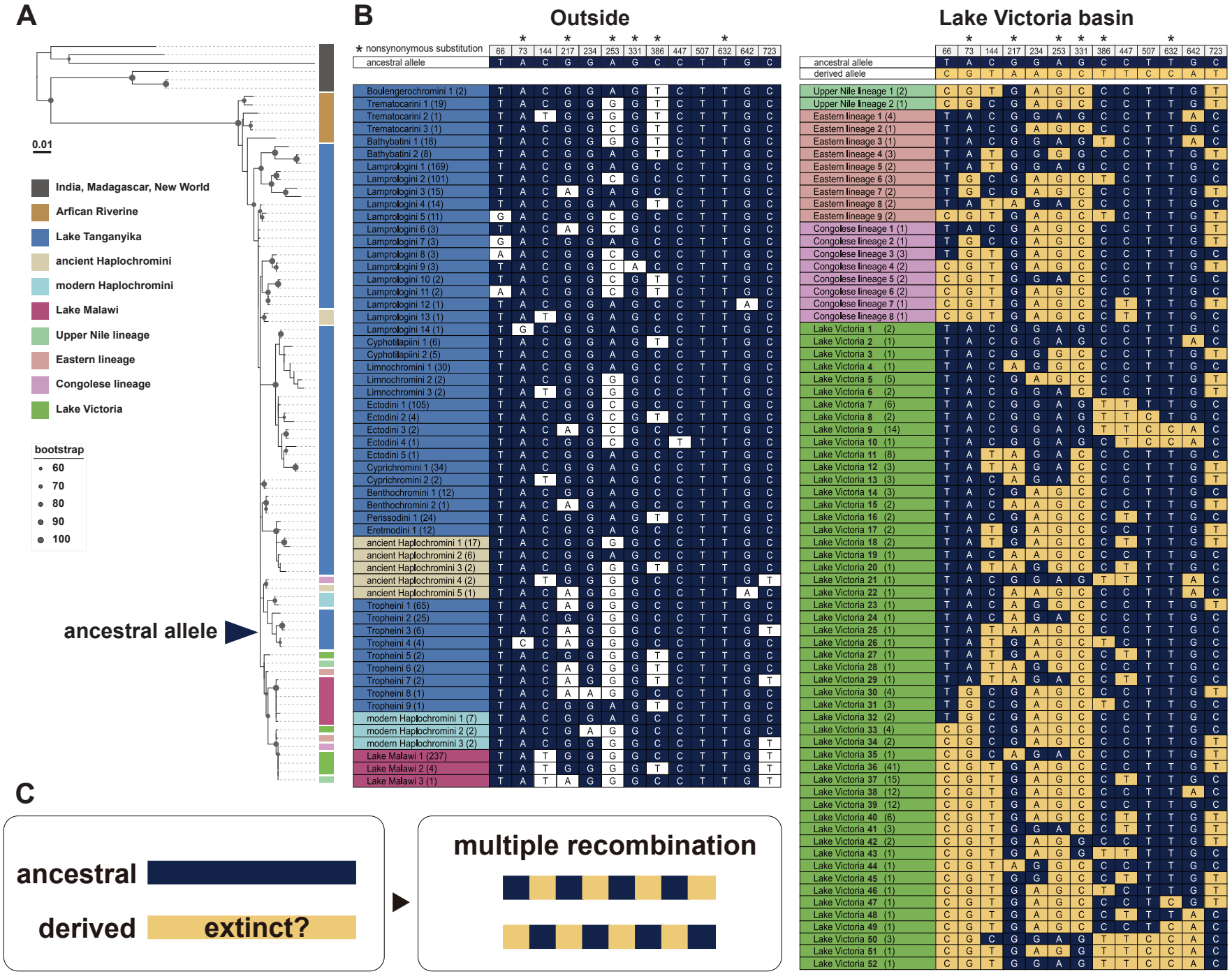

A
0.01
India, Madagascar, New World
Arfican Riverine
Lake Tanganyika
ancient Haplochromini
modern Haplochromini
Lake Malawi
Upper Nile lineage
Eastern lineage
Congolese lineage
Lake Victoria
bootstrap
60
70
80
90
100
ancestral allele
B
Outside
* nonsynonymous substitution
66 73 144 217 234 253 331 386 447 507 632 642 723
ancestral allele T A C G G A G C C T T G C
Boulengerochromini 1 (2) T A C G G A G T C T T G C
Trematocarini 1 (19) T A C G G G G T C T T G C
Trematocarini 2 (1) T A T G G G G T C T T G C
Trematocarini 3 (1) T A C G G C G T C T T G C
Bathybatini 1 (18) T A C G G G G T C T T G C
Bathybatini 2 (8) T A C G G A G T C T T G C
Lamprologini 1 (169) T A C G G A G C C T T G C
Lamprologini 2 (101) T A C G G C G C C T T G C
Lamprologini 3 (15) T A C A G A G C C T T G C
Lamprologini 4 (14) T A C G G A G T C T T G C
Lamprologini 5 (11) G A C G G C G C C T T G C
Lamprologini 6 (3) T A C A G C G C C T T G C
Lamprologini 7 (3) G A C G G A G C C T T G C
Lamprologini 8 (3) A A C G G C G C C T T G C
Lamprologini 9 (3) T A C G G C A C C T T G C
Lamprologini 10 (2) T A C G G C G T C T T G C
Lamprologini 11 (2) A A C G G C G T C T T G C
Lamprologini 12 (1) T A C G G A G C C T T A C
Lamprologini 13 (1) T A T G G A G C C T T G C
Lamprologini 14 (1) T G C G G A G C C T T G C
Cyphotilapiini 1 (6) T A C G G A G T C T T G C
Cyphotilapiini 2 (5) T A C G G A G C C T T G C
Limnochromini 1 (30) T A C G G A G C C T T G C
Limnochromini 2 (2) T A C G G G G C C T T G C
Limnochromini 3 (2) T A T G G G G C C T T G C
Ectodini 1 (105) T A C G G C G C C T T G C
Ectodini 2 (4) T A C G G C G T C T T G C
Ectodini 3 (2) T A C A G C G C C T T G C
Ectodini 4 (1) T A C G G C G C T T T G C
Ectodini 5 (1) T A C G G A G C C T T G C
Cyprichromini 1 (34) T A C G G A G C C T T G C
Cyprichromini 2 (2) T A T G G A G C C T T G C
Benthochromini 1 (12) T A C G G A G C C T T G C
Benthochromini 2 (1) T A C A G A G C C T T G C
Perissodini 1 (24) T A C G G A G T C T T G C
Eretmodini 1 (12) T A C G G A G C C T T G C
ancient Haplochromini 1 (17) T A C G G G G C C T T G C
ancient Haplochromini 2 (6) T A C G G A G C C T T G C
ancient Haplochromini 3 (2) T A C G G G G T C T T G C
ancient Haplochromini 4 (2) T A T G G G G C C T T G T
ancient Haplochromini 5 (1) T A C A G G G C C T T A C
Tropheini 1 (65) T A C A G G G C C T T G C
Tropheini 2 (25) T A C G G G G C C T T G C
Tropheini 3 (6) T A C A G G G C C T T G T
Tropheini 4 (4) T C C A G G G C C T T G C
Tropheini 5 (2) T A C G G G G T C T T G C
Tropheini 6 (2) T A C A G G G T C T T G C
Tropheini 7 (2) T A C A G G G T C T T G T
Tropheini 8 (1) T A C A A G G C C T T G C
Tropheini 9 (1) T A C G G A G T C T T G C
modern Haplochromini 1 (7) T A C G G A G C C T T G C
modern Haplochromini 2 (2) T A C G A G G C C T T G C
modern Haplochromini 3 (2) T A C G G G G C C T T G T
Lake Malawi 1 (237) T A T G G G G C C T T G T
Lake Malawi 2 (4) T A T G G G G T C T T G T
Lake Malawi 3 (1) T A T A G G G C C T T G T
Lake Victoria basin
66 73 144 217 234 253 331 386 447 507 632 642 723
ancestral allele T A C G G A G C C T T G C
derived allele C G T A A G C T T C C A T
Upper Nile lineage 1 (2) C G T G A G C C C T T G T
Upper Nile lineage 2 (1) C G C G A G C C C T T G T
Eastern lineage 1 (4) T A C G G A G C C T T A C
Eastern lineage 2 (1) T A C G A G C C C T T G C
Eastern lineage 3 (1) T A C G G A G T C T T A C
Eastern lineage 4 (3) T A T G G G G C C T T G T
Eastern lineage 5 (2) T A T G G A G C C T T G C
Eastern lineage 6 (3) T G C G A G C T C T T G C
Eastern lineage 7 (2) T G C G A G C C C T T G T
Eastern lineage 8 (2) T A T A G A C C C T T G C
Eastern lineage 9 (2) C G T G A G C T C T T G T
Congolese lineage 1 (1) T A C G A G C C C T T G T
Congolese lineage 2 (1) T G C G A G C C C T T G T
Congolese lineage 3 (3) T G T G A G C C C T T G C
Congolese lineage 4 (2) C G T G A G C C C T T G T
Congolese lineage 5 (2) C G T G G A C C C T T G C
Congolese lineage 6 (2) C G T G A G C C C T T G C
Congolese lineage 7 (1) C G T G A G C C T T T G T
Congolese lineage 8 (1) C G T G A G C C T T T G C
Lake Victoria 1 (2) T A C G G A G C C T T G C
Lake Victoria 2 (1) T A C G G A G C C T T A C
Lake Victoria 3 (1) T A C G G G C C C T T G T
Lake Victoria 4 (1) T A C A G G C C C T T G C
Lake Victoria 5 (5) T A C G A G C C C T T G T
Lake Victoria 6 (2) T A C G G A C C C T T G T
Lake Victoria 7 (6) T A C G G A G T T T T G C
Lake Victoria 8 (2) T A C G G A G T T C T G C
Lake Victoria 9 (14) T A C G G A G T T C C A C
Lake Victoria 10 (1) T A C G G A G C T C C A C
Lake Victoria 11 (8) T A T A G A C C C T T G C
Lake Victoria 12 (3) T A T A G A C C C T T G T
Lake Victoria 13 (3) T A C A G A C C C T T G T
Lake Victoria 14 (3) T A C G A G C C C T T G C
Lake Victoria 15 (2) T A C A A G C C C T T G T
Lake Victoria 16 (2) T A C G A G C C T T T G C
Lake Victoria 17 (2) T A T G A G C C C T T G T
Lake Victoria 18 (2) T A T G A G C C T T T G T
Lake Victoria 19 (1) T A C A A G C C C T T G C
Lake Victoria 20 (1) T A T A G G C C T T T G C
Lake Victoria 21 (1) T A C G G A G T T T T A C
Lake Victoria 22 (1) T A C A A G C C C T T A C
Lake Victoria 23 (1) T A C A G G C C C T T G T
Lake Victoria 24 (1) T A C A G A C C C T T G C
Lake Victoria 25 (1) T A T A A G C C C T T G C
Lake Victoria 26 (1) T A T G A G C T C T T G C
Lake Victoria 27 (1) T A T G A G C C T T T G C
Lake Victoria 28 (1) T A T A G G C C C T T G C
Lake Victoria 29 (1) T A T A G A C C T T T G C
Lake Victoria 30 (4) T G C G A G C C C T T G T
Lake Victoria 31 (3) T G C G A G C T C T T G C
Lake Victoria 32 (2) T G C G A G C C C T T G C
Lake Victoria 33 (4) C G C G A G C C C T T G C
Lake Victoria 34 (2) C G C G A G C C C T T G T
Lake Victoria 35 (1) C G C A G A C C C T T G T
Lake Victoria 36 (41) C G T G A G C C C T T G T
Lake Victoria 37 (15) C G T G A G C C T T T G C
Lake Victoria 38 (12) C G T G A G C C C T T A C
Lake Victoria 39 (12) C G T G A G C C C T T G C
Lake Victoria 40 (6) C G T G A G C C T T T G T
Lake Victoria 41 (3) C G T G G A C C T T T G T
Lake Victoria 42 (2) C G T G A G G C C T T G T
Lake Victoria 43 (1) C G T G A G G T T T T G C
Lake Victoria 44 (1) C G T A G G C C C T T G C
Lake Victoria 45 (1) C G T G G G C C T T T G T
Lake Victoria 46 (1) C G T G A G C T C T T G T
Lake Victoria 47 (1) C G T G A G C C C T C G T
Lake Victoria 48 (1) C G T G A G C C T T T A C
Lake Victoria 49 (1) C G T G A G C C C T C A C
Lake Victoria 50 (3) C G C G G A G T T C C A C
Lake Victoria 51 (1) C G T G A G G T T C C A C
Lake Victoria 52 (1) C G T G G A G T T C C A C
C
ancestral
derived
extinct?
multiple recombination

Figure 3

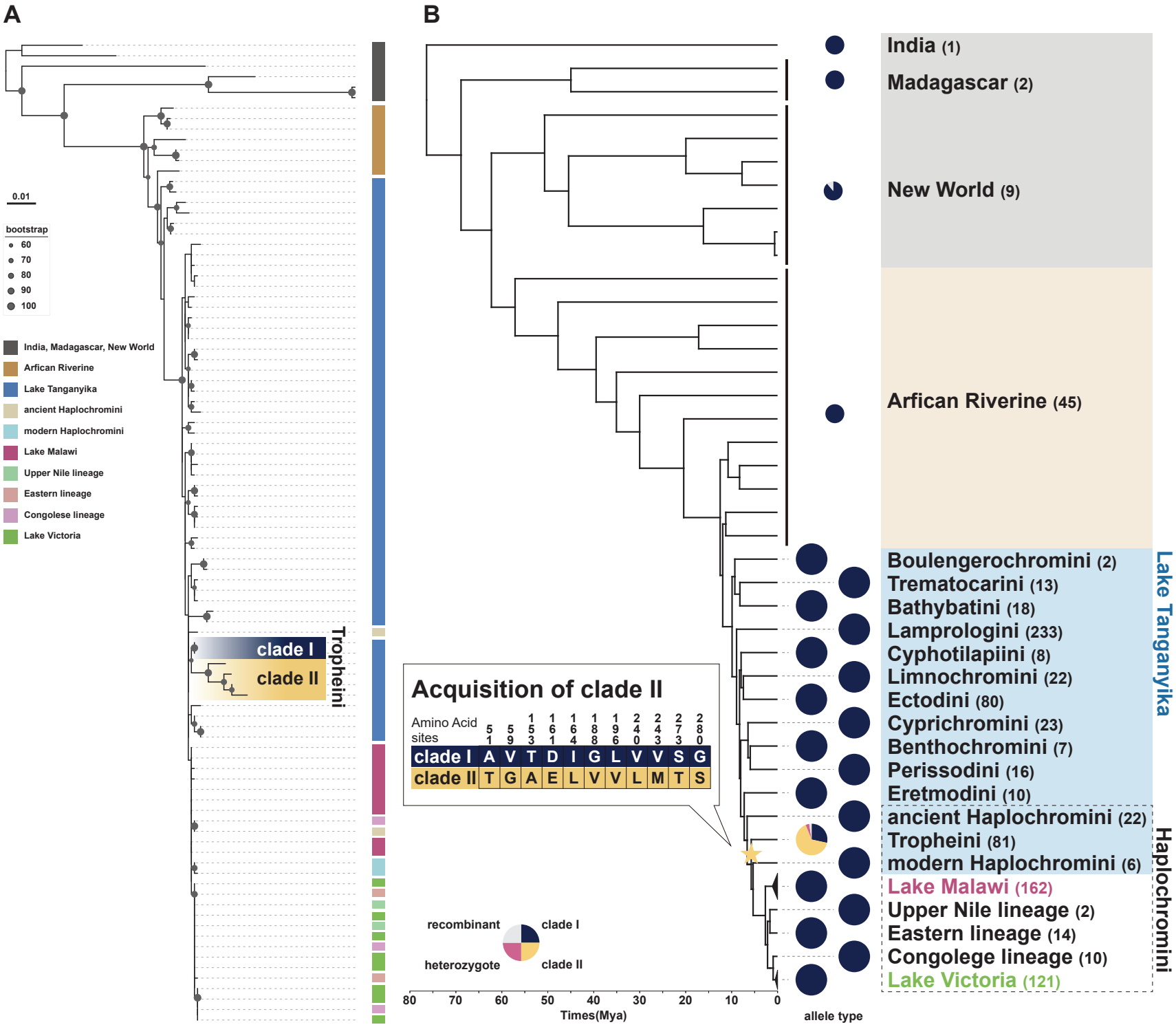

A
B
0.01
bootstrap
60
70
80
90
100
India, Madagascar, New World
Arfican Riverine
Lake Tanganyika
ancient Haplochromini
modern Haplochromini
Lake Malawi
Upper Nile lineage
Eastern lineage
Congolese lineage
Lake Victoria
clade I
clade II
Tropheini
Acquisition of clade II
Amino Acid sites
51 59 153 161 164 188 196 240 243 273 280
clade I A V T D I G L V V S G
clade II T G A E L V V L M T S
recombinant
clade I
heterozygote
clade II
80 70 60 50 40 30 20 10 0
Times(Mya)
allele type
India (1)
Madagascar (2)
New World (9)
Arfican Riverine (45)
Boulengerochromini (2)
Trematocarini (13)
Bathybatini (18)
Lamprologini (233)
Cyphotilapiini (8)
Limnochromini (22)
Ectodini (80)
Cyprichromini (23)
Benthochromini (7)
Perissodini (16)
Eretmodini (10)
Lake Tanganyika
ancient Haplochromini (22)
Tropheini (81)
modern Haplochromini (6)
Lake Malawi (162)
Upper Nile lineage (2)
Eastern lineage (14)
Congolege lineage (10)
Lake Victoria (121)
Haplochromini

Figure 4

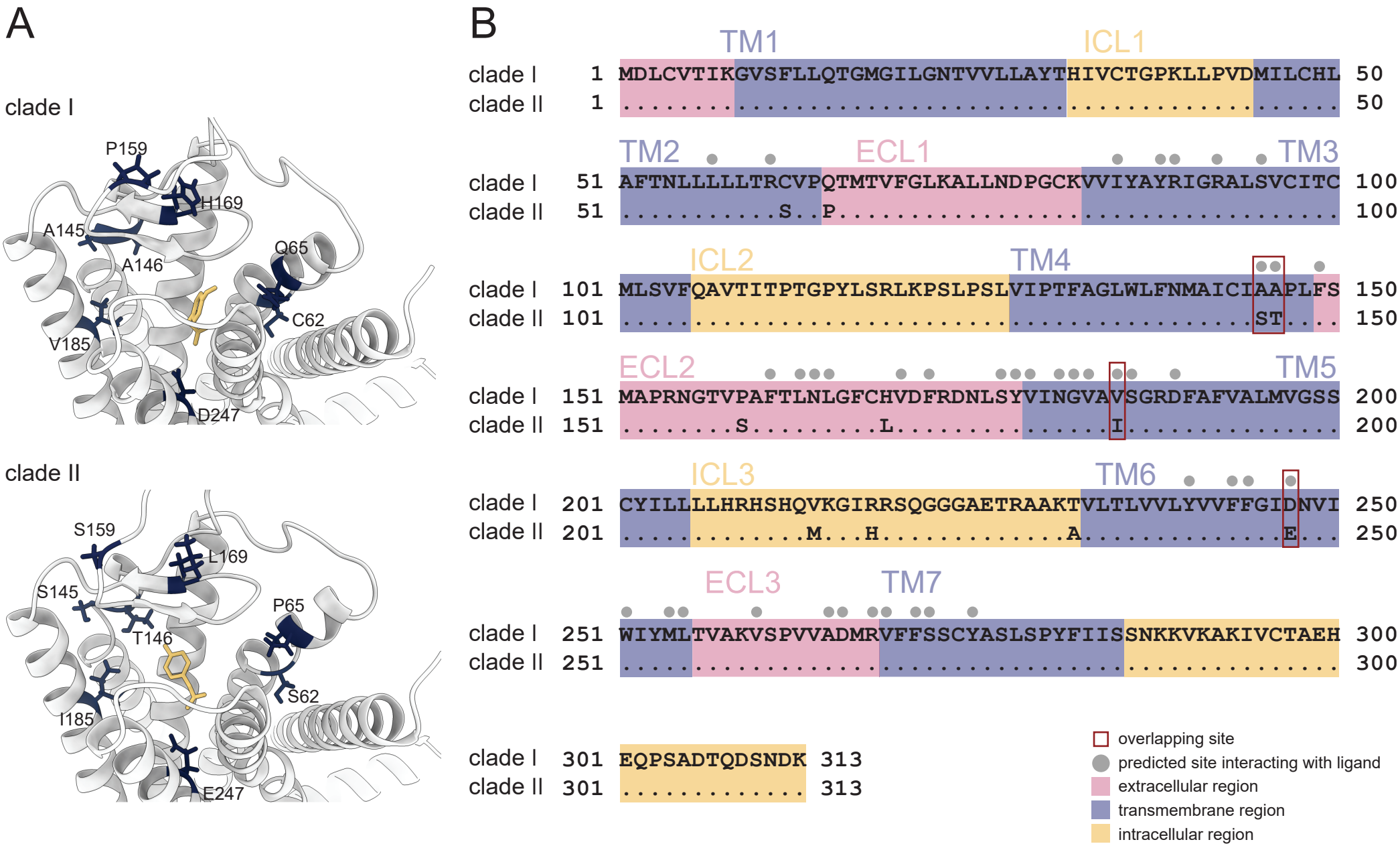

A
clade I
P159
H169
A145
A146
Q65
C62
V185
D247
clade II
S159
L169
S145
T146
P65
S62
I185
E247
B
TM1
ICL1
clade I 1 MDLCVTIKGVSFLLQTGMGILGNTVVLLAYTHIVCTGPKLLPVDMILCHL 50
clade II 1 50
TM2
ECL1
TM3
clade I 51 AFTNLLLLLTRCVPQTMTVFGLKALLNDPGCKVVIYAYRIGRALSVCITC 100
clade II 51 .....S..P..... 100
ICL2
TM4
clade I 101 MLSVFQAVTITPTGPYLSRLKPSLPSLVIPTFAGLWLFNMAICIAAPLFS 150
clade II 101 .....ST.... 150
ECL2
TM5
clade I 151 MAPRNGTVPAFTLNLGFCHVDFRDNLSYVINGVAVSGRDFAFVALMVGSS 200
clade II 151 .....S.....L.....I..... 200
ICL3
TM6
clade I 201 CYILLLLHRHSHQVKGIRRSQGGGAETRAAKTVLTLVVLYVVFFGIDNVI 250
clade II 201 .....M...H.....A.....E... 250
ECL3
TM7
clade I 251 WIYMLTVAKVSPVVADMRVFFSSCYASLSPYFIISSNKKVKAKIVCTAEH 300
clade II 251 300
clade I 301 EQPSADTQDSNDK 313
clade II 301 313
overlapping site
predicted site interacting with ligand
extracellular region
transmembrane region
intracellular region

Figure 5

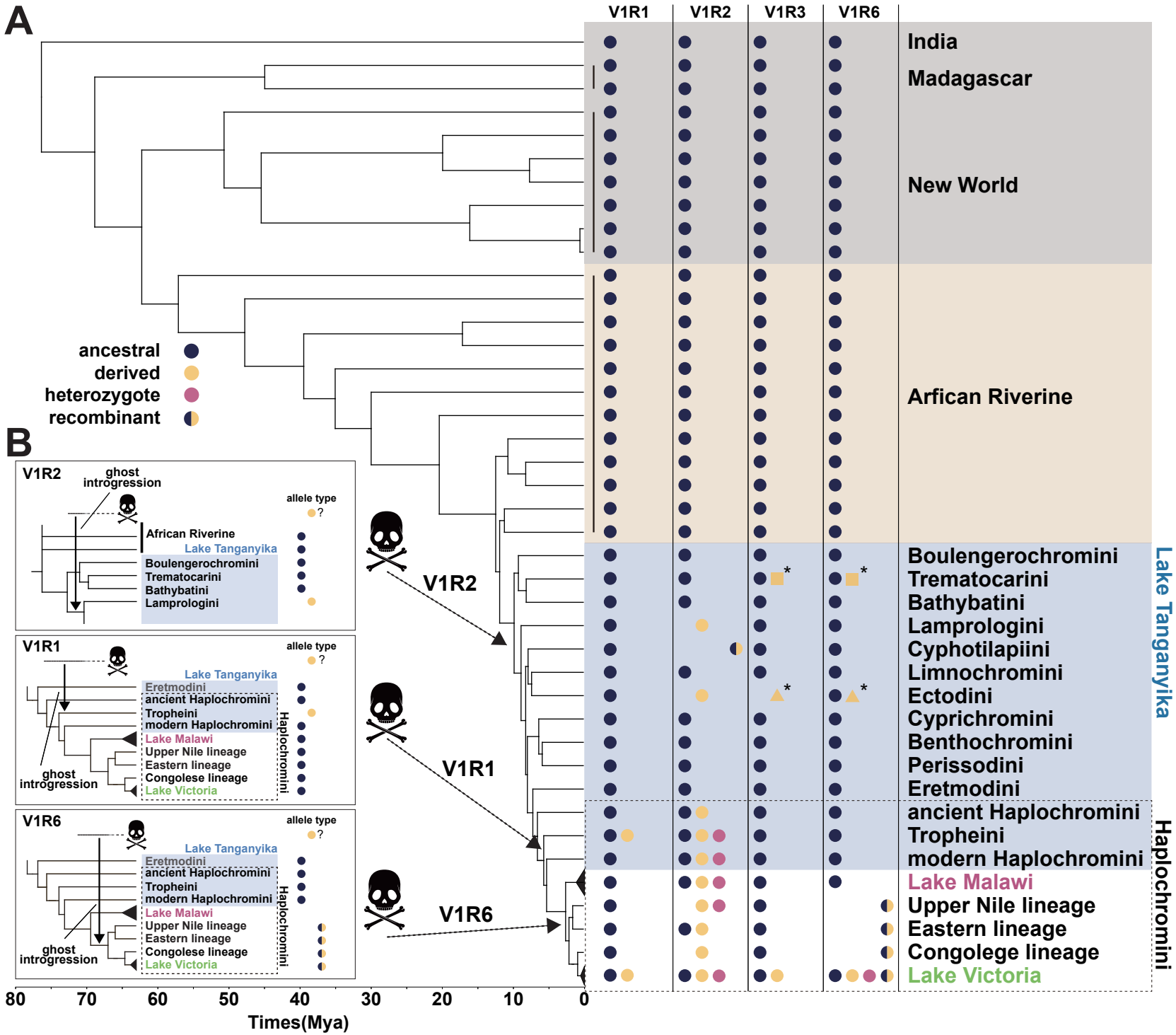

A
V1R1
V1R2
V1R3
V1R6
India
Madagascar
New World
Arfican Riverine
Lake Tanganyika
Boulengerochromini
Trematocarini
Bathybatini
Lamprologini
Cyphotilapiini
Limnochromini
Ectodini
Cyprichromini
Benthochromini
Perissodini
Eretmodini
Haplochromini
ancient Haplochromini
Tropheini
modern Haplochromini
Lake Malawi
Upper Nile lineage
Eastern lineage
Congolege lineage
Lake Victoria
ancestral
derived
heterozygote
recombinant
B
V1R2
ghost introgression
allele type
African Riverine
Lake Tanganyika
Boulengerochromini
Trematocarini
Bathybatini
Lamprologini
V1R1
Eretmodini
ancient Haplochromini
Tropheini
modern Haplochromini
Lake Malawi
Upper Nile lineage
Eastern lineage
Congolese lineage
Lake Victoria
V1R6
80
70
60
50
40
30
20
10
0
Times(Mya)

Figure 6

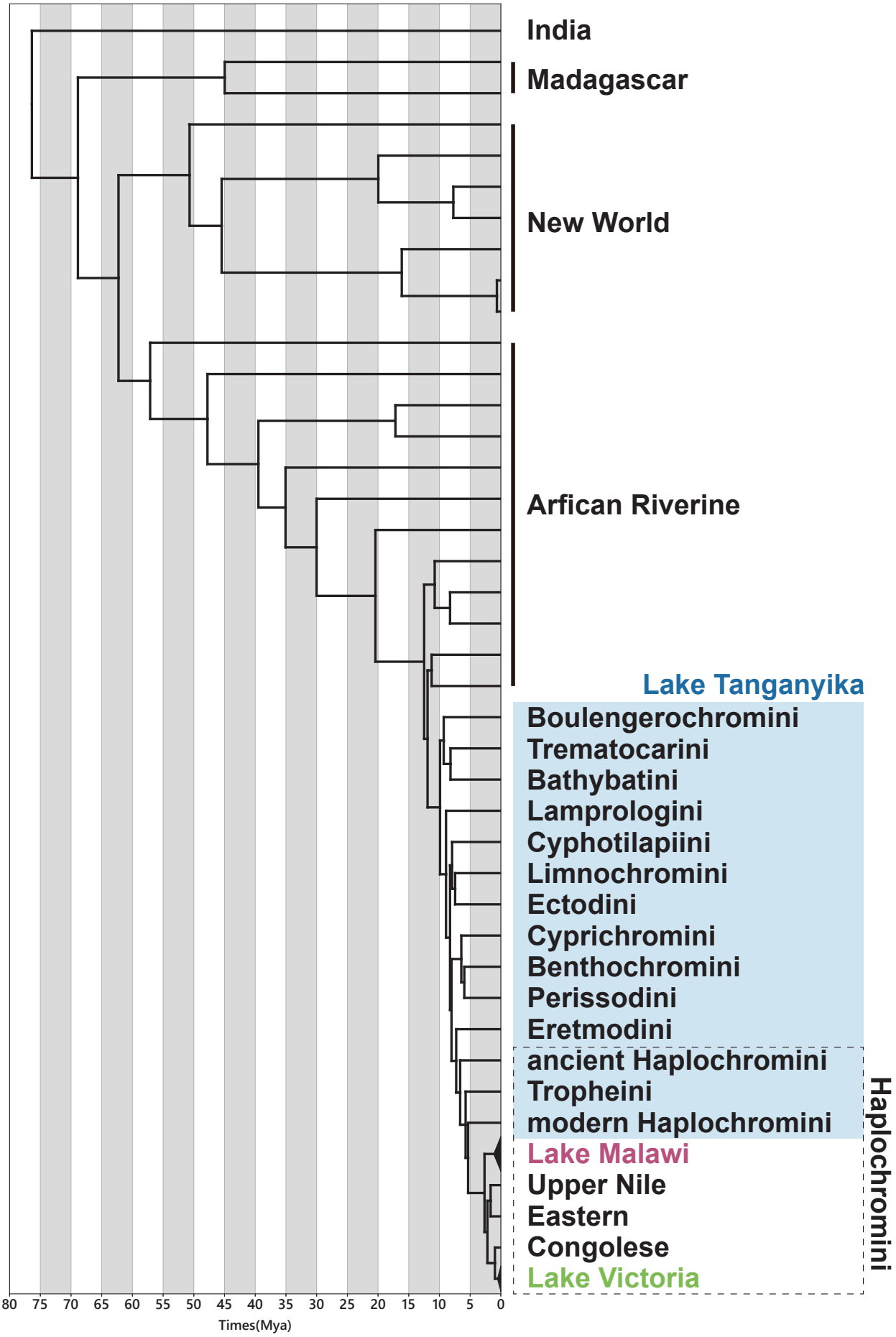
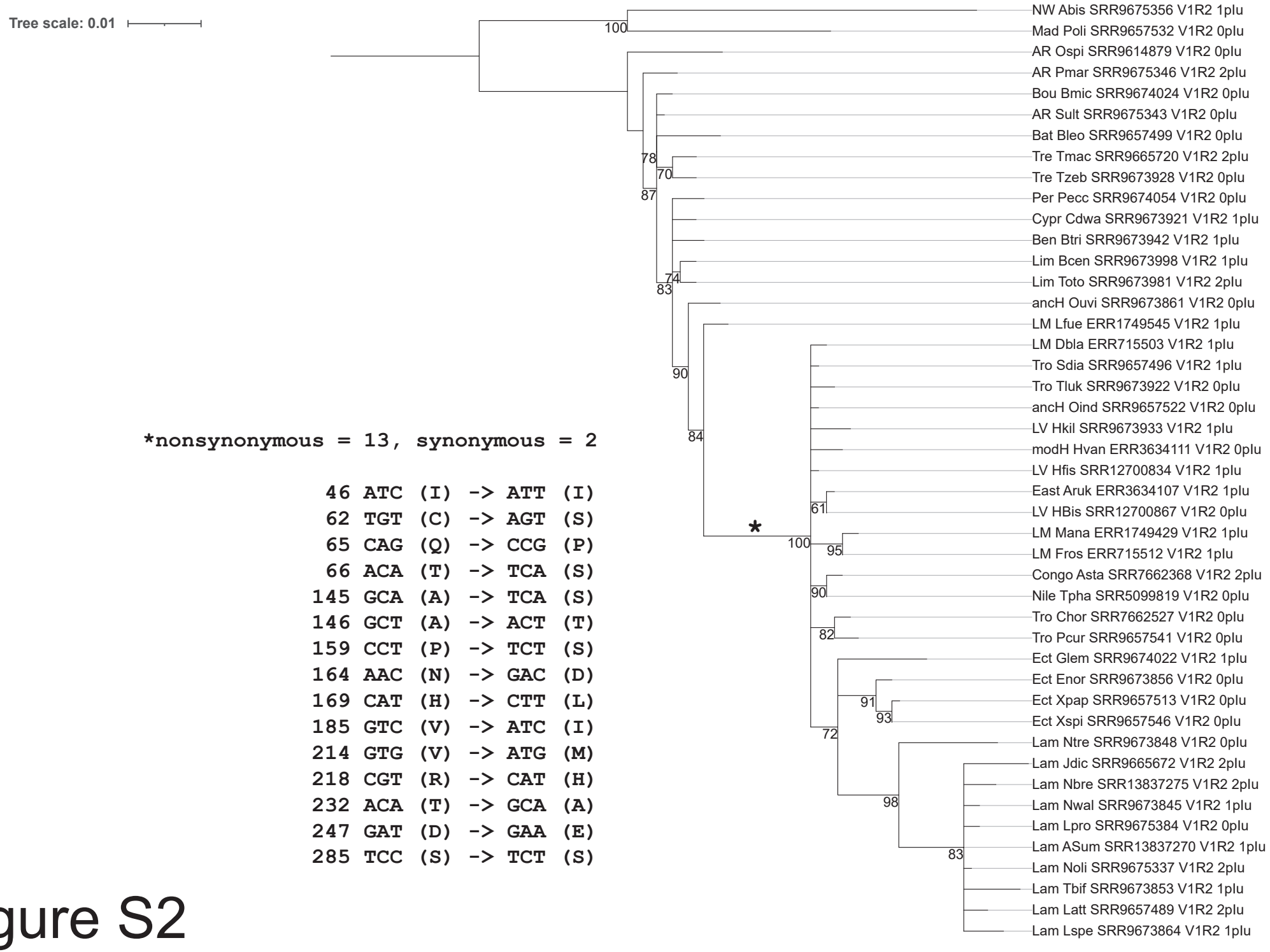

India
Madagascar
New World
Arfican Riverine
Lake Tanganyika
Boulengerochromini
Trematocarini
Bathybatini
Lamprologini
Cyphotilapiini
Limnochromini
Ectodini
Cyprichromini
Benthochromini
Perissodini
Eretmodini
ancient Haplochromini
Tropheini
modern Haplochromini
Lake Malawi
Upper Nile
Eastern
Congolese
Lake Victoria
Haplochromini
80 75 70 65 60 55 50 45 40 35 30 25 20 15 10 5 0
Times(Mya)

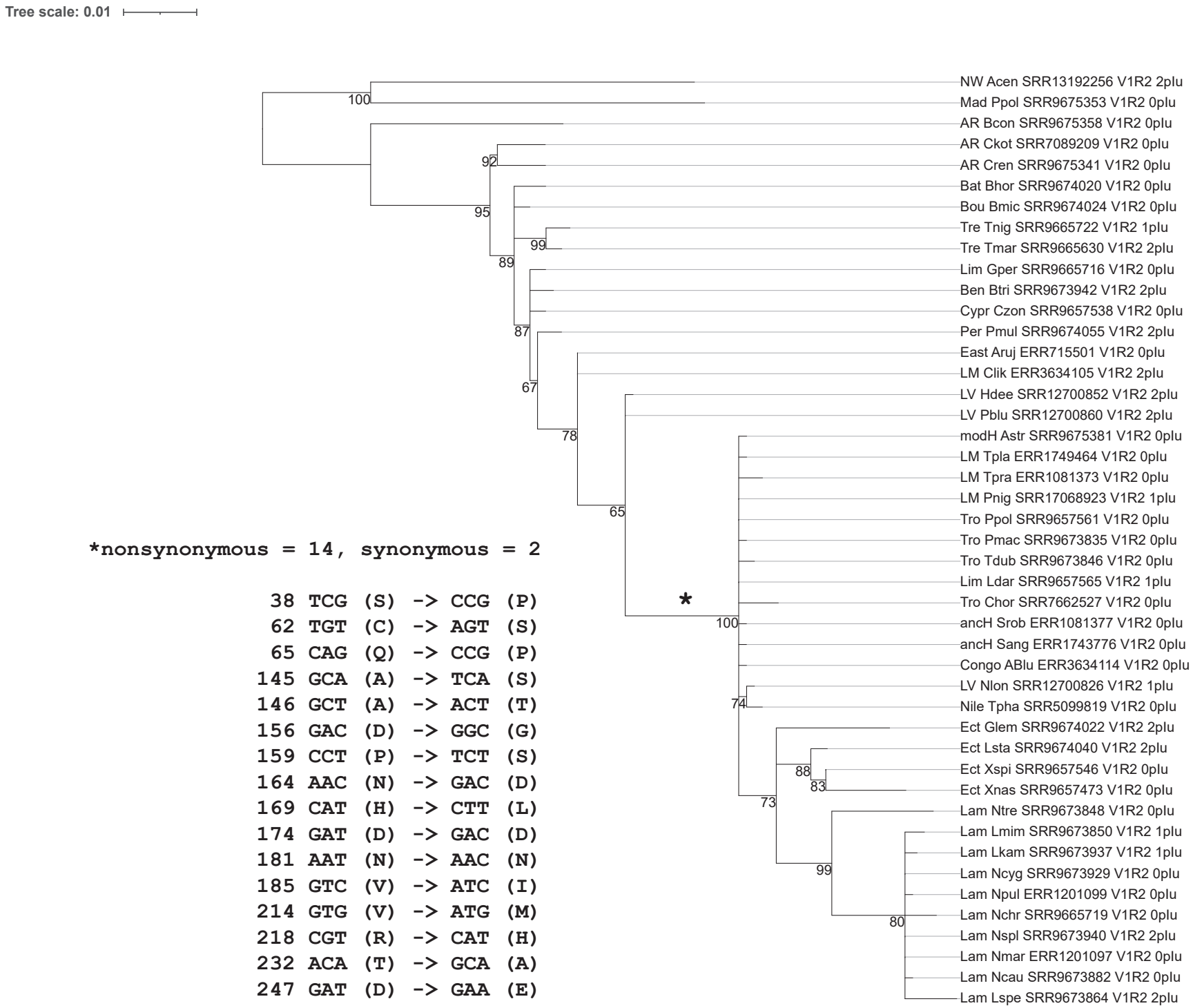
Figure S1



*nonsynonymous = 13, synonymous = 2

46	ATC	(I)	->	ATT	(I)
62	TGT	(C)	->	AGT	(S)
65	CAG	(Q)	->	CCG	(P)
66	ACA	(T)	->	TCA	(S)
145	GCA	(A)	->	TCA	(S)
146	GCT	(A)	->	ACT	(T)
159	CCT	(P)	->	TCT	(S)
164	AAC	(N)	->	GAC	(D)
169	CAT	(H)	->	CTT	(L)
185	GTC	(V)	->	ATC	(I)
214	GTG	(V)	->	ATG	(M)
218	CGT	(R)	->	CAT	(H)
232	ACA	(T)	->	GCA	(A)
247	GAT	(D)	->	GAA	(E)
285	TCC	(S)	->	TCT	(S)

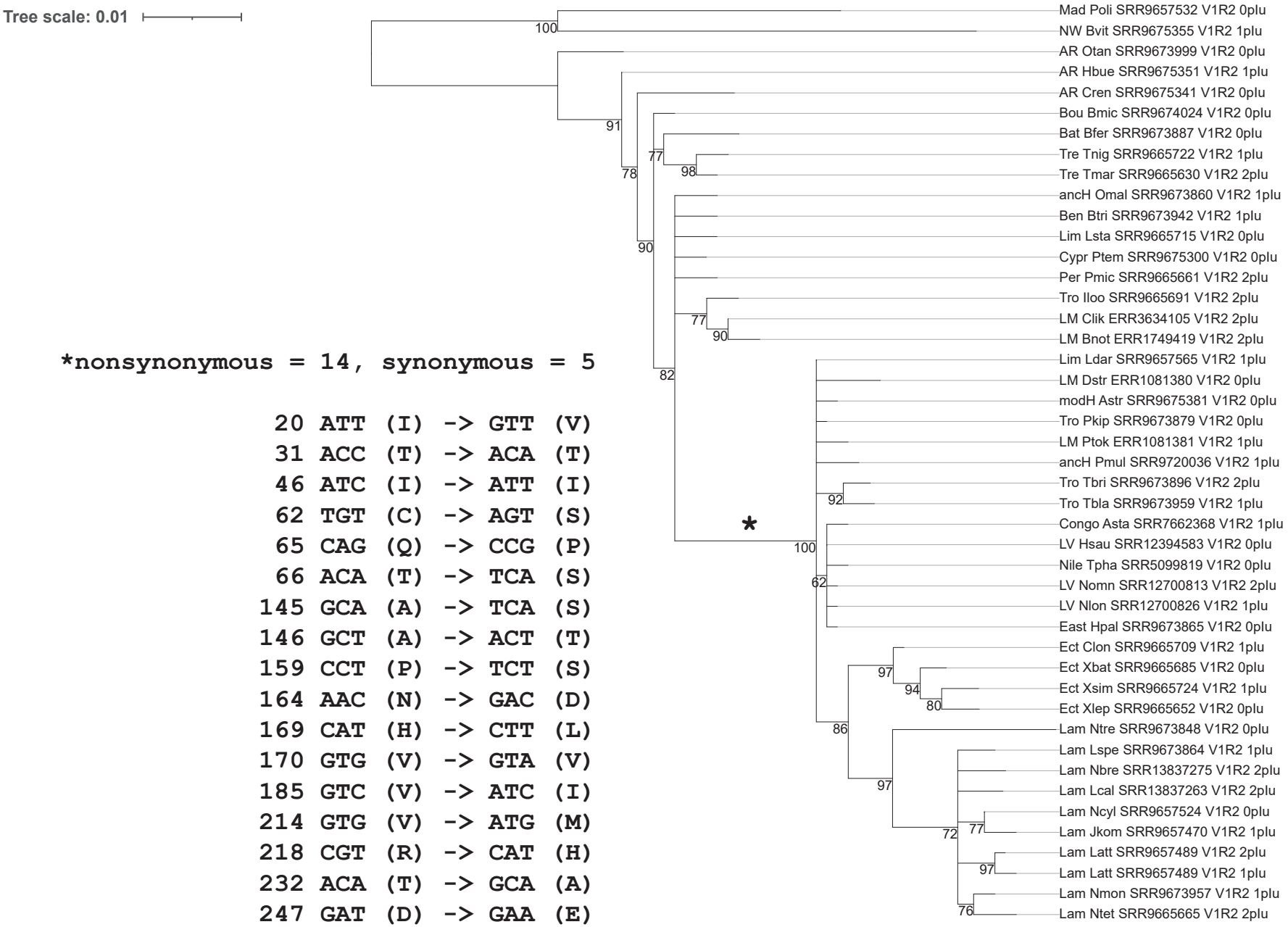
Figure S2



*nonsynonymous = 14, synonymous = 2

38	TCG	(S)	->	CCG	(P)
62	TGT	(C)	->	AGT	(S)
65	CAG	(Q)	->	CCG	(P)
145	GCA	(A)	->	TCA	(S)
146	GCT	(A)	->	ACT	(T)
156	GAC	(D)	->	GGC	(G)
159	CCT	(P)	->	TCT	(S)
164	AAC	(N)	->	GAC	(D)
169	CAT	(H)	->	CTT	(L)
174	GAT	(D)	->	GAC	(D)
181	AAT	(N)	->	AAC	(N)
185	GTC	(V)	->	ATC	(I)
214	GTG	(V)	->	ATG	(M)
218	CGT	(R)	->	CAT	(H)
232	ACA	(T)	->	GCA	(A)
247	GAT	(D)	->	GAA	(E)

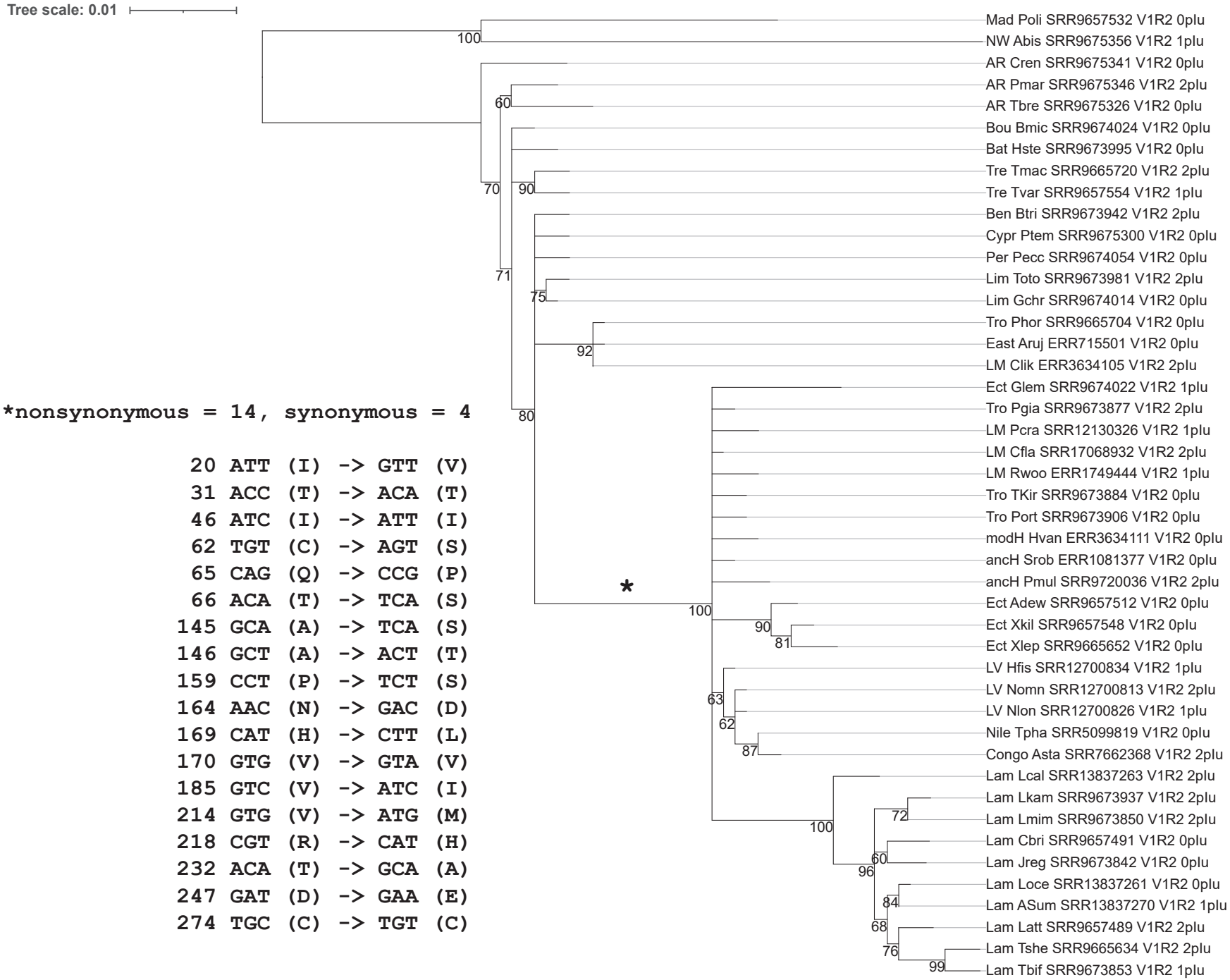
Figure S3



*nonsynonymous = 14, synonymous = 5

20	ATT	(I)	->	GTT	(V)
31	ACC	(T)	->	ACA	(T)
46	ATC	(I)	->	ATT	(I)
62	TGT	(C)	->	AGT	(S)
65	CAG	(Q)	->	CCG	(P)
66	ACA	(T)	->	TCA	(S)
145	GCA	(A)	->	TCA	(S)
146	GCT	(A)	->	ACT	(T)
159	CCT	(P)	->	TCT	(S)
164	AAC	(N)	->	GAC	(D)
169	CAT	(H)	->	CTT	(L)
170	GTG	(V)	->	GTA	(V)
185	GTC	(V)	->	ATC	(I)
214	GTG	(V)	->	ATG	(M)
218	CGT	(R)	->	CAT	(H)
232	ACA	(T)	->	GCA	(A)
247	GAT	(D)	->	GAA	(E)
274	TGC	(C)	->	TGT	(C)
285	TCC	(S)	->	TCT	(S)

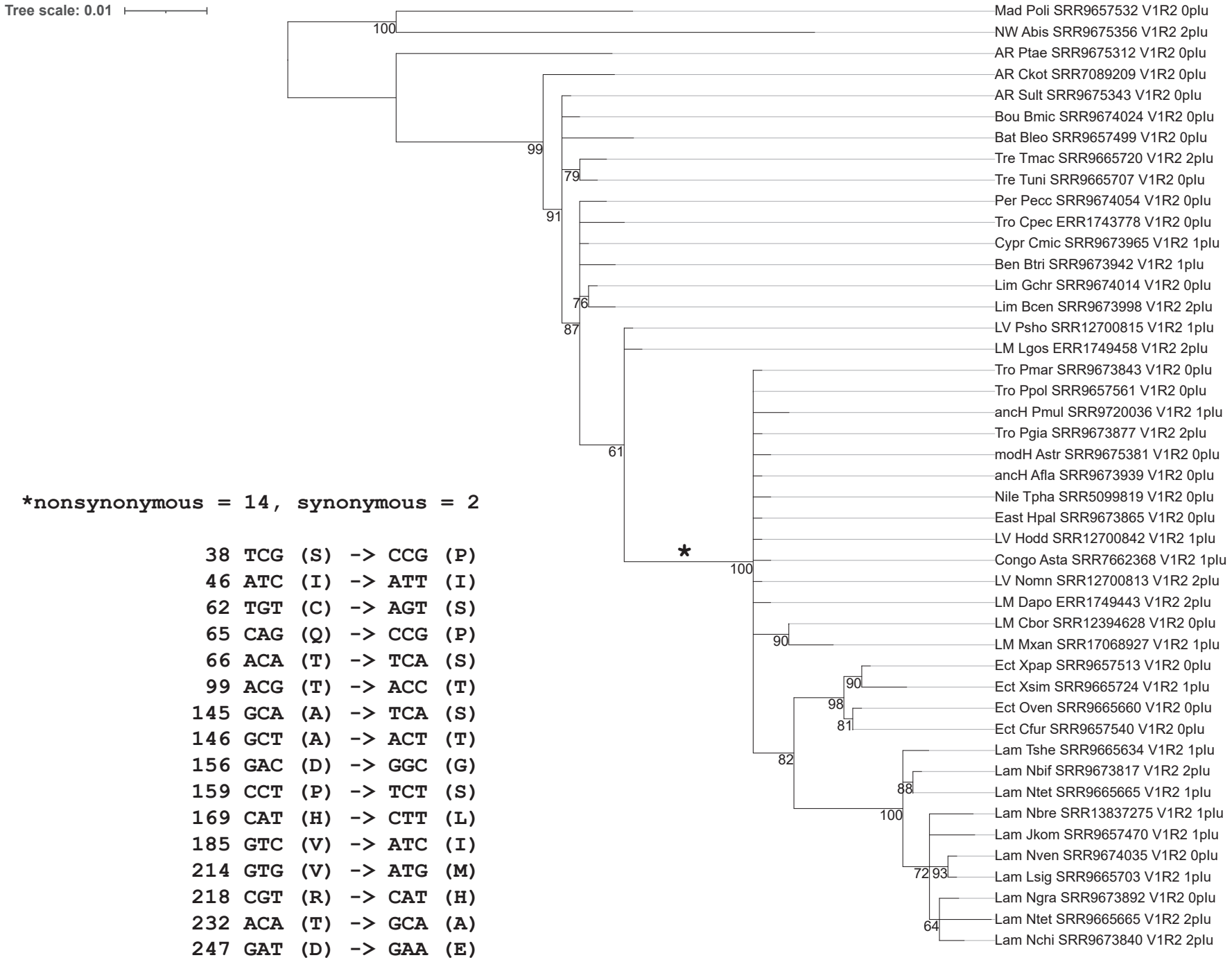
Figure S4



*nonsynonymous = 14, synonymous = 4

20	ATT	(I)	->	GTT	(V)
31	ACC	(T)	->	ACA	(T)
46	ATC	(I)	->	ATT	(I)
62	TGT	(C)	->	AGT	(S)
65	CAG	(Q)	->	CCG	(P)
66	ACA	(T)	->	TCA	(S)
145	GCA	(A)	->	TCA	(S)
146	GCT	(A)	->	ACT	(T)
159	CCT	(P)	->	TCT	(S)
164	AAC	(N)	->	GAC	(D)
169	CAT	(H)	->	CTT	(L)
170	GTG	(V)	->	GTA	(V)
185	GTC	(V)	->	ATC	(I)
214	GTG	(V)	->	ATG	(M)
218	CGT	(R)	->	CAT	(H)
232	ACA	(T)	->	GCA	(A)
247	GAT	(D)	->	GAA	(E)
274	TGC	(C)	->	TGT	(C)

Figure S5



***nonsynonymous = 14, synonymous = 2**

38	TCG	(S)	->	CCG	(P)
46	ATC	(I)	->	ATT	(I)
62	TGT	(C)	->	AGT	(S)
65	CAG	(Q)	->	CCG	(P)
66	ACA	(T)	->	TCA	(S)
99	ACG	(T)	->	ACC	(T)
145	GCA	(A)	->	TCA	(S)
146	GCT	(A)	->	ACT	(T)
156	GAC	(D)	->	GGC	(G)
159	CCT	(P)	->	TCT	(S)
169	CAT	(H)	->	CTT	(L)
185	GTC	(V)	->	ATC	(I)
214	GTG	(V)	->	ATG	(M)
218	CGT	(R)	->	CAT	(H)
232	ACA	(T)	->	GCA	(A)
247	GAT	(D)	->	GAA	(E)

Figure S6

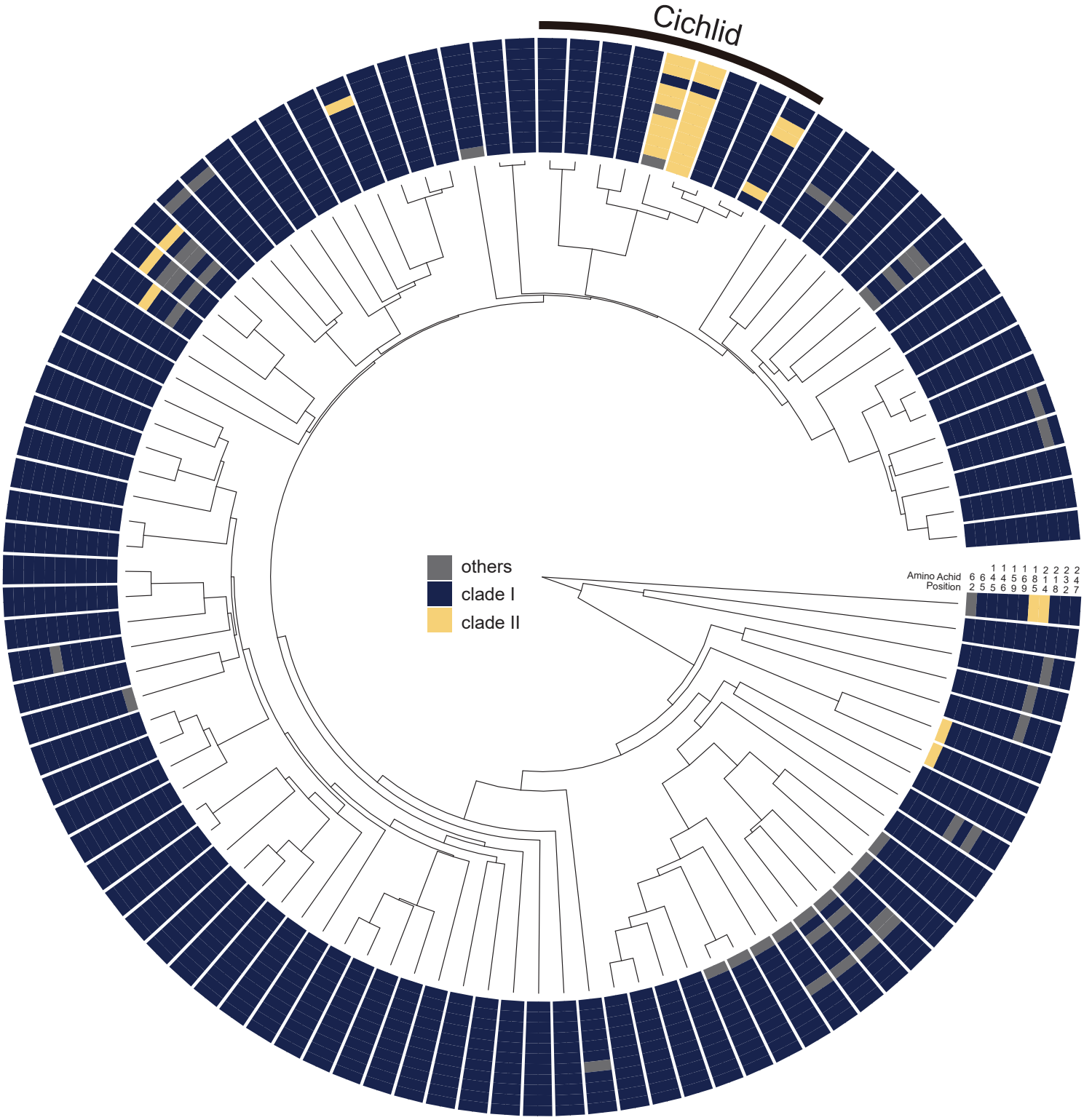

Cichlid
others
clade I
clade II
Amino Achid
Position
6 6 1 1 1 1 2 2 2 2
2 5 4 5 6 9 5 4 8 2 7
1 4 1 3 4

Figure S7

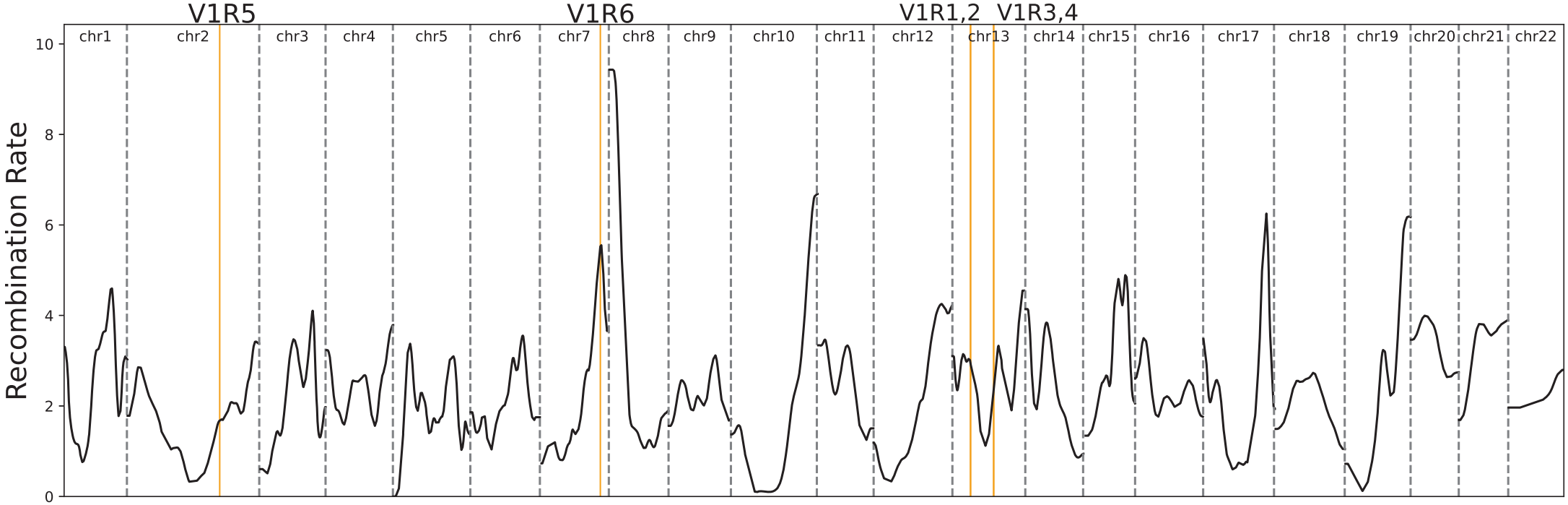

V1R5
V1R6
V1R1,2
V1R3,4
chr1
chr2
chr3
chr4
chr5
chr6
chr7
chr8
chr9
chr10
chr11
chr12
chr13
chr14
chr15
chr16
chr17
chr18
chr19
chr20
chr21
chr22
Recombination Rate
0
2
4
6
8
10

Figure S8

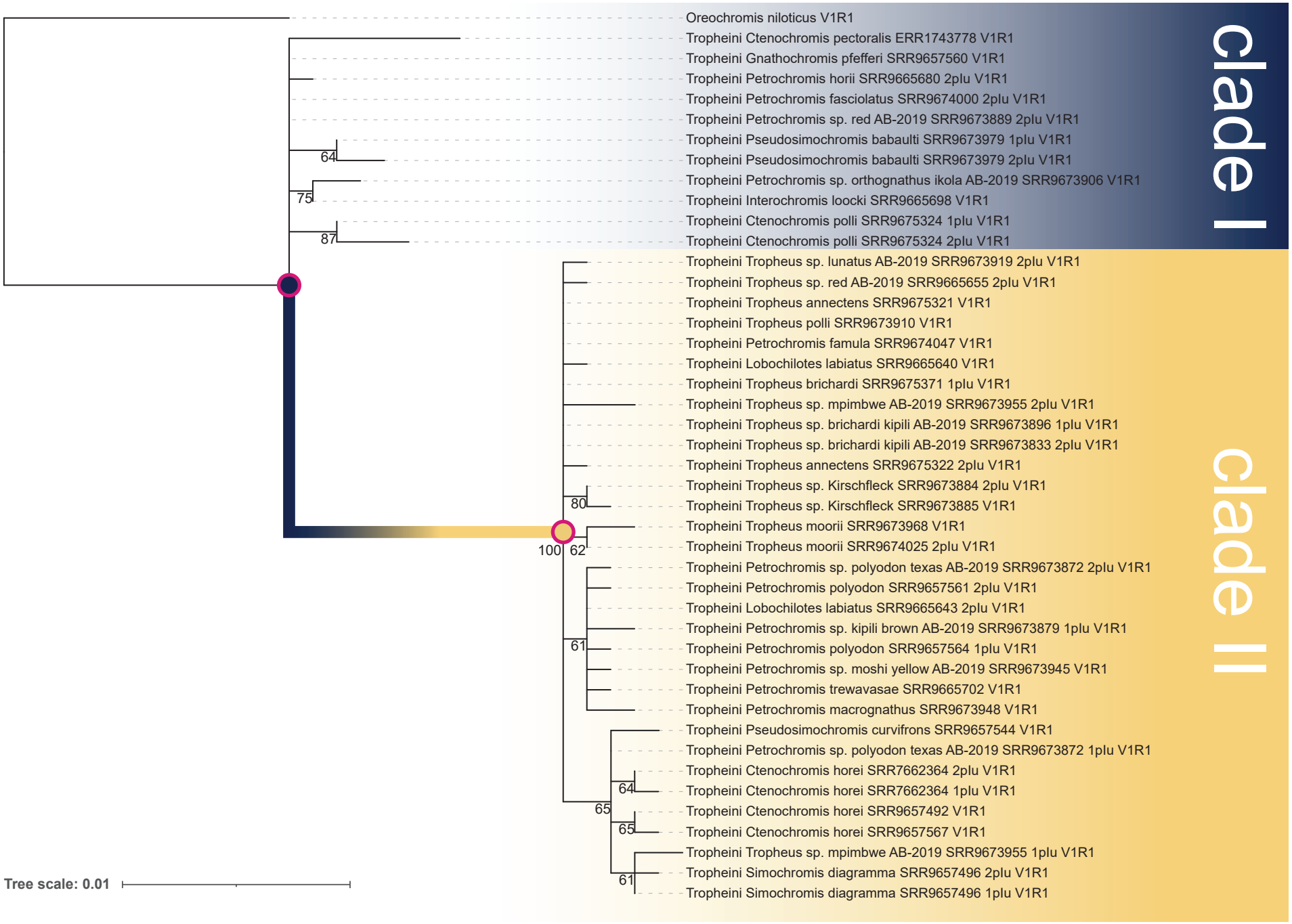

Oreochromis niloticus V1R1
Tropheini Ctenochromis pectoralis ERR1743778 V1R1
Tropheini Gnathochromis pfefferi SRR9657560 V1R1
Tropheini Petrochromis horii SRR9665680 2plu V1R1
Tropheini Petrochromis fasciolatus SRR9674000 2plu V1R1
Tropheini Petrochromis sp. red AB-2019 SRR9673889 2plu V1R1
Tropheini Pseudosimochromis babaulti SRR9673979 1plu V1R1
Tropheini Pseudosimochromis babaulti SRR9673979 2plu V1R1
Tropheini Petrochromis sp. orthognathus ikola AB-2019 SRR9673906 V1R1
Tropheini Interochromis loocki SRR9665698 V1R1
Tropheini Ctenochromis polli SRR9675324 1plu V1R1
Tropheini Ctenochromis polli SRR9675324 2plu V1R1
Tropheini Tropheus sp. lunatus AB-2019 SRR9673919 2plu V1R1
Tropheini Tropheus sp. red AB-2019 SRR9665655 2plu V1R1
Tropheini Tropheus annectens SRR9675321 V1R1
Tropheini Tropheus polli SRR9673910 V1R1
Tropheini Petrochromis famula SRR9674047 V1R1
Tropheini Lobochilotes labiatus SRR9665640 V1R1
Tropheini Tropheus brichardi SRR9675371 1plu V1R1
Tropheini Tropheus sp. mpimbwe AB-2019 SRR9673955 2plu V1R1
Tropheini Tropheus sp. brichardi kipili AB-2019 SRR9673896 1plu V1R1
Tropheini Tropheus sp. brichardi kipili AB-2019 SRR9673833 2plu V1R1
Tropheini Tropheus annectens SRR9675322 2plu V1R1
Tropheini Tropheus sp. Kirschfleck SRR9673884 2plu V1R1
Tropheini Tropheus sp. Kirschfleck SRR9673885 V1R1
Tropheini Tropheus moorii SRR9673968 V1R1
Tropheini Tropheus moorii SRR9674025 2plu V1R1
Tropheini Petrochromis sp. polyodon texas AB-2019 SRR9673872 2plu V1R1
Tropheini Petrochromis polyodon SRR9657561 2plu V1R1
Tropheini Lobochilotes labiatus SRR9665643 2plu V1R1
Tropheini Petrochromis sp. kipili brown AB-2019 SRR9673879 1plu V1R1
Tropheini Petrochromis polyodon SRR9657564 1plu V1R1
Tropheini Petrochromis sp. moshi yellow AB-2019 SRR9673945 V1R1
Tropheini Petrochromis trewavasae SRR9665702 V1R1
Tropheini Petrochromis macrognathus SRR9673948 V1R1
Tropheini Pseudosimochromis curvifrons SRR9657544 V1R1
Tropheini Petrochromis sp. polyodon texas AB-2019 SRR9673872 1plu V1R1
Tropheini Ctenochromis horei SRR7662364 2plu V1R1
Tropheini Ctenochromis horei SRR7662364 1plu V1R1
Tropheini Ctenochromis horei SRR9657492 V1R1
Tropheini Ctenochromis horei SRR9657567 V1R1
Tropheini Tropheus sp. mpimbwe AB-2019 SRR9673955 1plu V1R1
Tropheini Simochromis diagramma SRR9657496 2plu V1R1
Tropheini Simochromis diagramma SRR9657496 1plu V1R1
64
75
87
80
100
62
61
64
65
65
61
clade I
clade II
Tree scale: 0.01

Figure S9

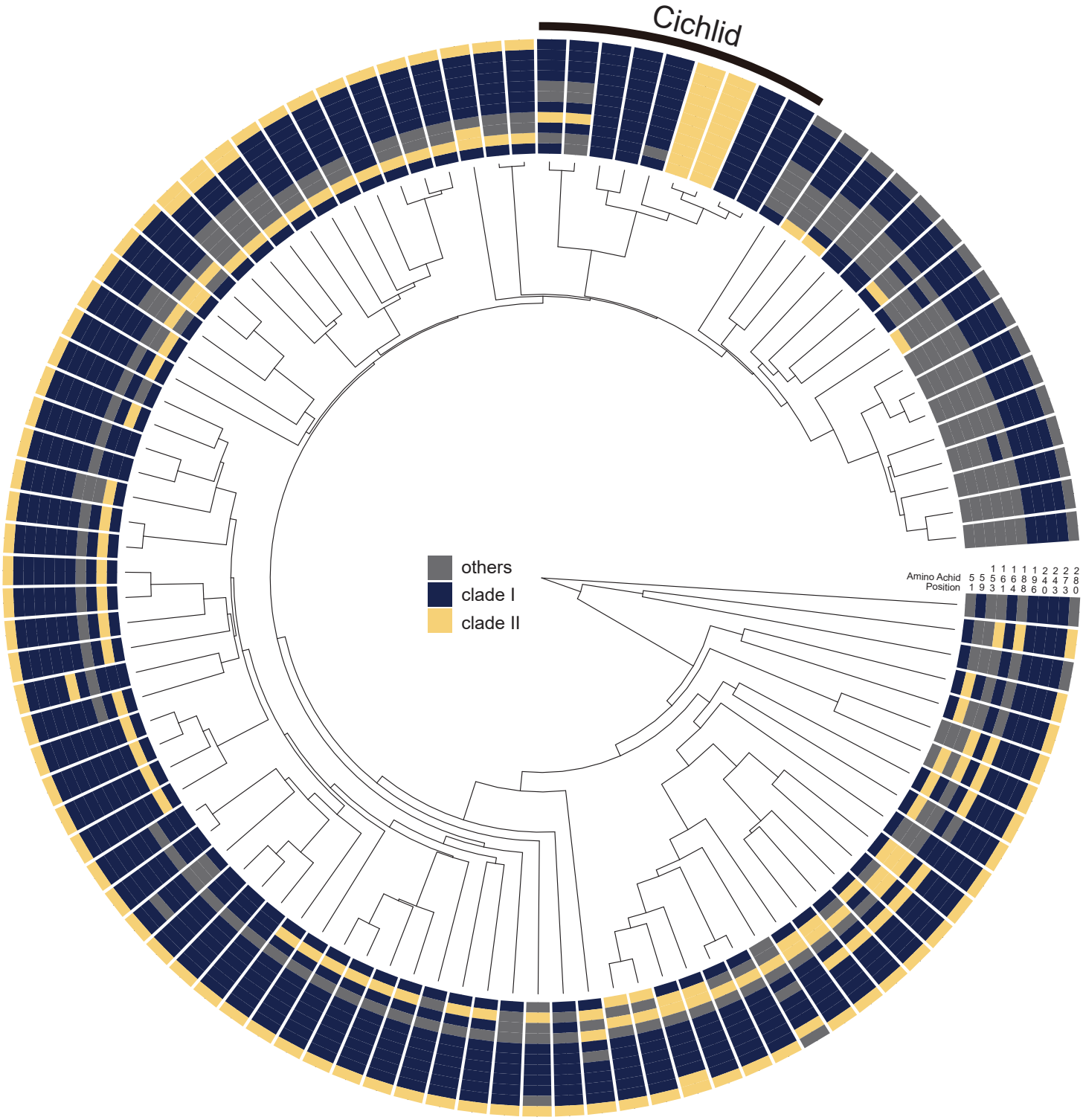

Cichlid
others
clade I
clade II
Amino Achid Position
51
59
153
161
164
188
196
240
243
273
280

Figure S10

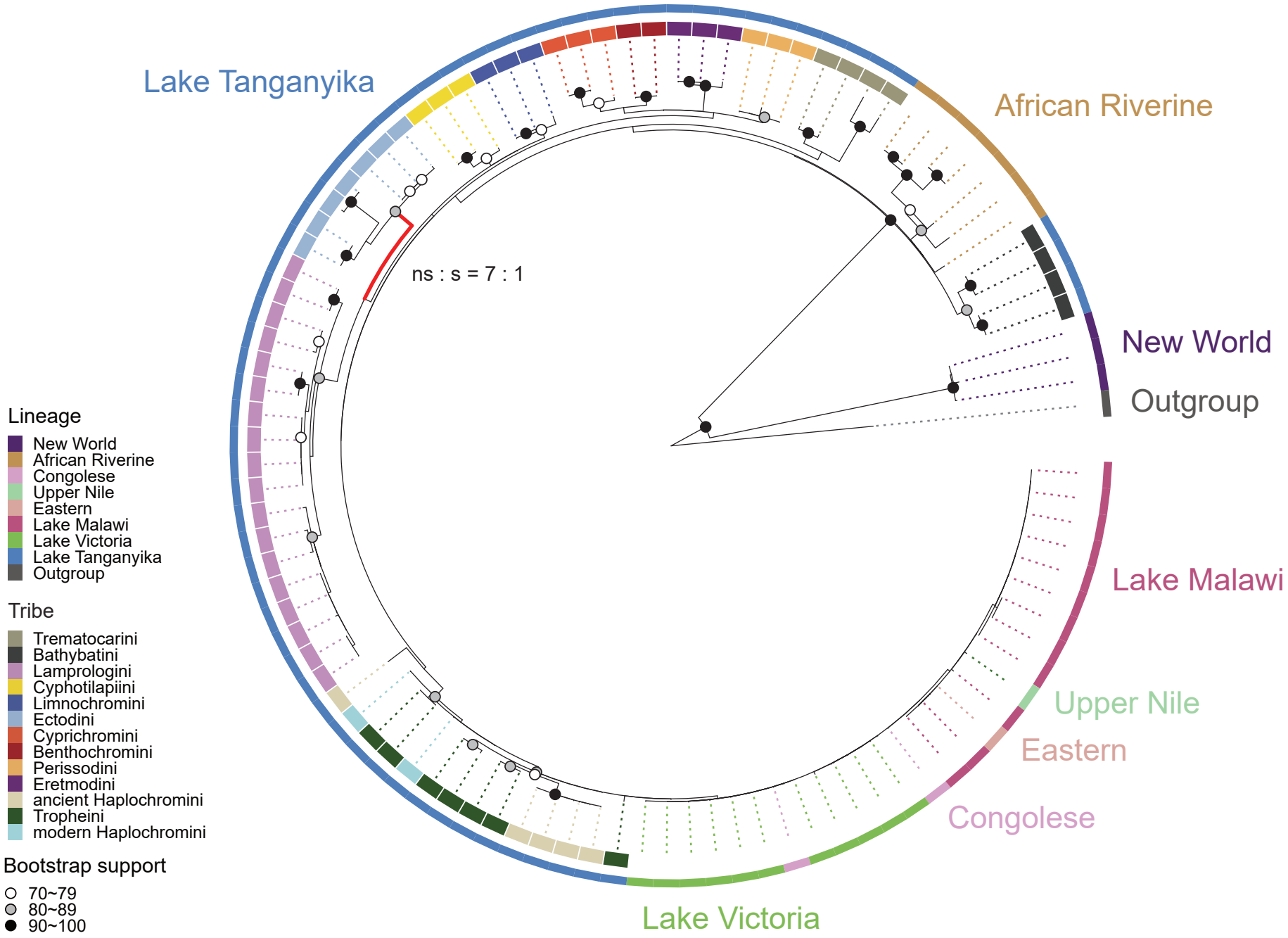

Lake Tanganyika
African Riverine
New World
Outgroup
Lake Malawi
Upper Nile
Eastern
Congolese
Lake Victoria
ns : s = 7 : 1
Lineage
New World
African Riverine
Congolese
Upper Nile
Eastern
Lake Malawi
Lake Victoria
Lake Tanganyika
Outgroup
Tribe
Trematocarini
Bathybatini
Lamprologini
Cyphotilapiini
Limnochromini
Ectodini
Cyprichromini
Benthochromini
Perissodini
Eretmodini
ancient Haplochromini
Tropheini
modern Haplochromini
Bootstrap support
70~79
80~89
90~100

Figure S11

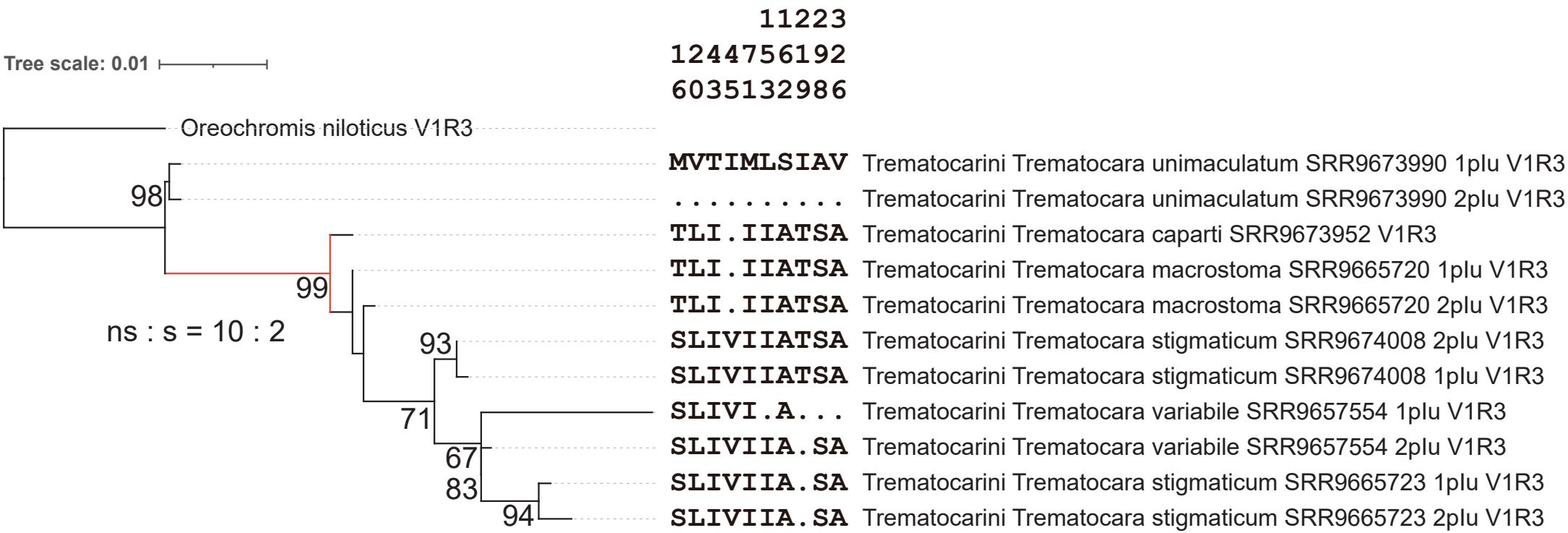

Tree scale: 0.01
11223
1244756192
6035132986
Oreochromis niloticus V1R3
98
99
ns : s = 10 : 2
93
71
67
83
94
MVTIMLSIAV Trematocarini Trematocara unimaculatum SRR9673990 1plu V1R3
..... Trematocarini Trematocara unimaculatum SRR9673990 2plu V1R3
TLI.IIATSA Trematocarini Trematocara caparti SRR9673952 V1R3
TLI.IIATSA Trematocarini Trematocara macrostoma SRR9665720 1plu V1R3
TLI.IIATSA Trematocarini Trematocara macrostoma SRR9665720 2plu V1R3
SLIVIIATSA Trematocarini Trematocara stigmaticum SRR9674008 2plu V1R3
SLIVIIATSA Trematocarini Trematocara stigmaticum SRR9674008 1plu V1R3
SLIVI.A... Trematocarini Trematocara variabile SRR9657554 1plu V1R3
SLIVIIA.SA Trematocarini Trematocara variabile SRR9657554 2plu V1R3
SLIVIIA.SA Trematocarini Trematocara stigmaticum SRR9665723 1plu V1R3
SLIVIIA.SA Trematocarini Trematocara stigmaticum SRR9665723 2plu V1R3

Figure S12

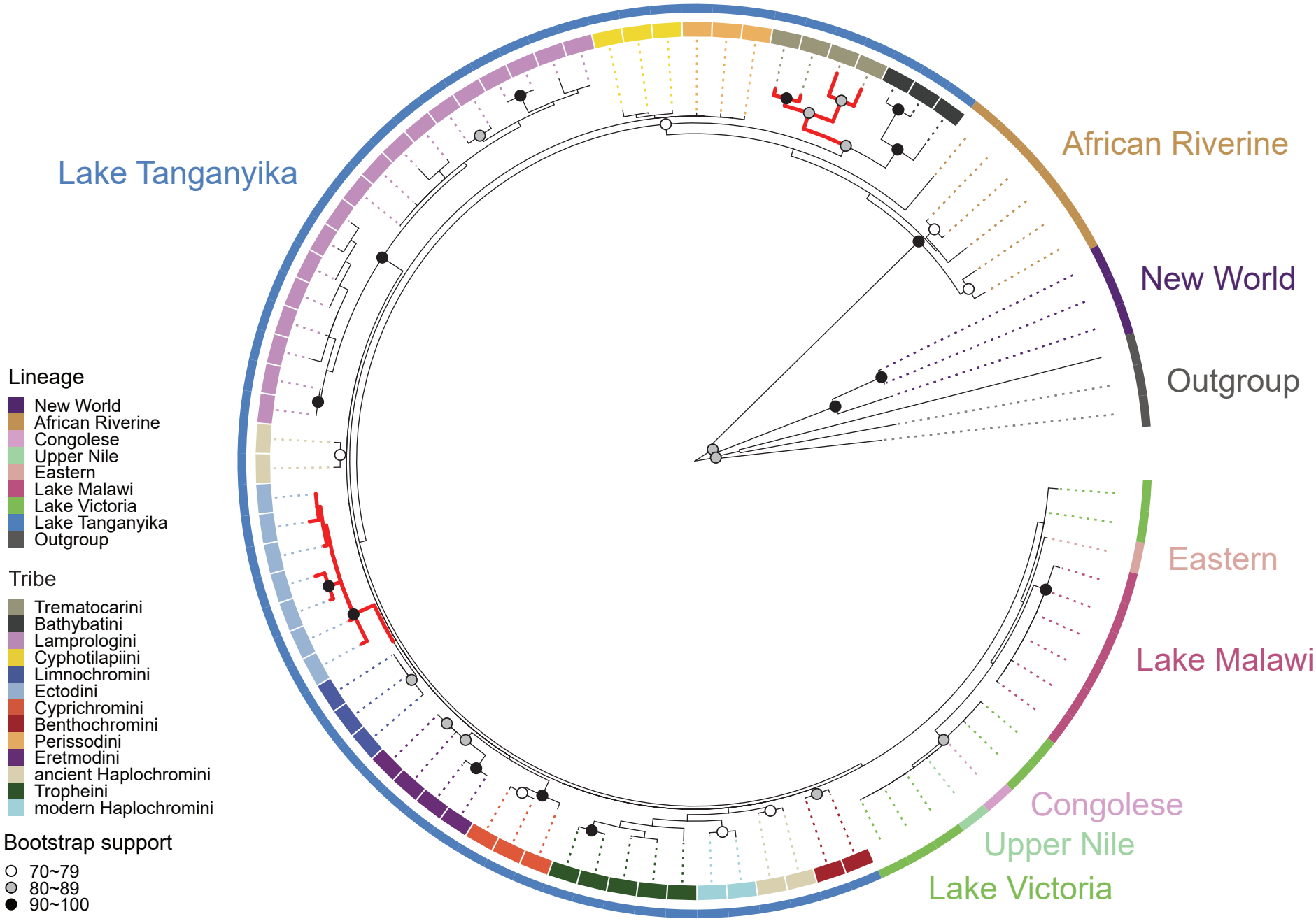

Lake Tanganyika
African Riverine
New World
Outgroup
Eastern
Lake Malawi
Congolese
Upper Nile
Lake Victoria
Lineage
New World
African Riverine
Congolese
Upper Nile
Eastern
Lake Malawi
Lake Victoria
Lake Tanganyika
Outgroup
Tribe
Trematocarini
Bathybatini
Lamprologini
Cyphotilapiini
Limnochromini
Ectodini
Cyprichromini
Benthochromini
Perissodini
Eretmodini
ancient Haplochromini
Tropheini
modern Haplochromini
Bootstrap support
70~79
80~89
90~100

Figure S13

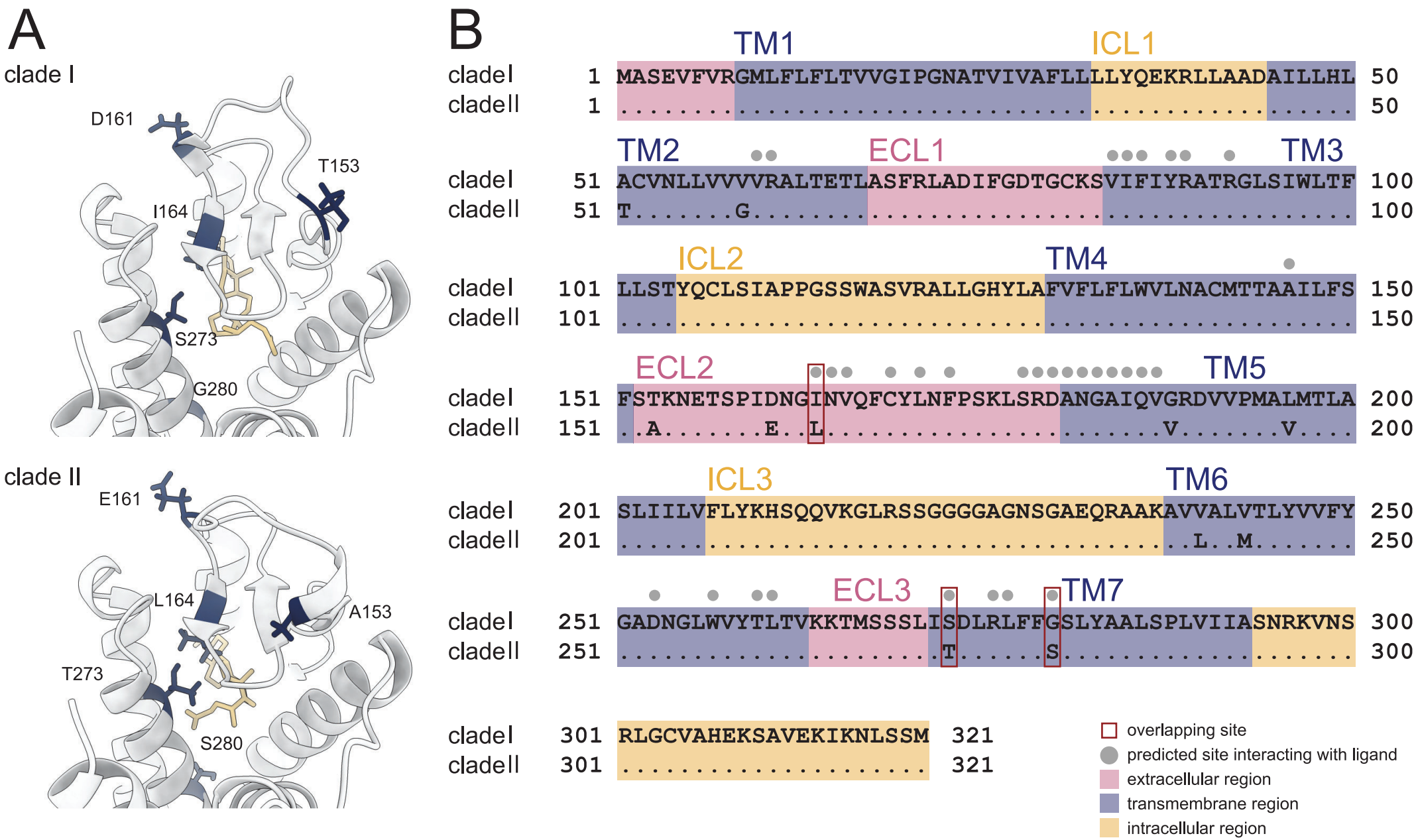

A
clade I
D161
T153
I164
S273
G280
clade II
E161
L164
A153
T273
S280
B
TM1
ICL1
cladeI 1 MASEVFVRGMLFLFLTVVGIPGNATVIVAFLLLYQEKRLLAADAILLHL 50
cladeII 1 50
TM2
ECL1
TM3
cladeI 51 ACVNLLVVVVRALTETLASFRLADIFGDTGCKSVIFIYRATRGLSIWLTF 100
cladeII 51 T.....G..... 100
ICL2
TM4
cladeI 101 LLSTYQCLSIAPPGSSWASVRALLGHYLAFVFLFLWVLNACMTTAAILFS 150
cladeII 101 150
ECL2
TM5
cladeI 151 FSTKNETSPIDNGINVQFCYLNFPSKLSRDANGAIQVGRDVVPMALMTLA 200
cladeII 151 ..A.....E..L.....V.....V.... 200
ICL3
TM6
cladeI 201 SLIILVFLYKHSQQVKGLRSSGGGGAGNSGAEQRAAKAVVALVTLYVVFY 250
cladeII 201 .....L..M..... 250
ECL3
TM7
cladeI 251 GADNGLWVYTLTVKKTMSSSLISDLRLFFGSLYAALSPLVIIASNRKVNS 300
cladeII 251 .....T.....S..... 300
cladeI 301 RLGCVAHEKSAVEKIKNLSSM 321
cladeII 301 321
overlapping site
predicted site interacting with ligand
extracellular region
transmembrane region
intracellular region

Figure S14

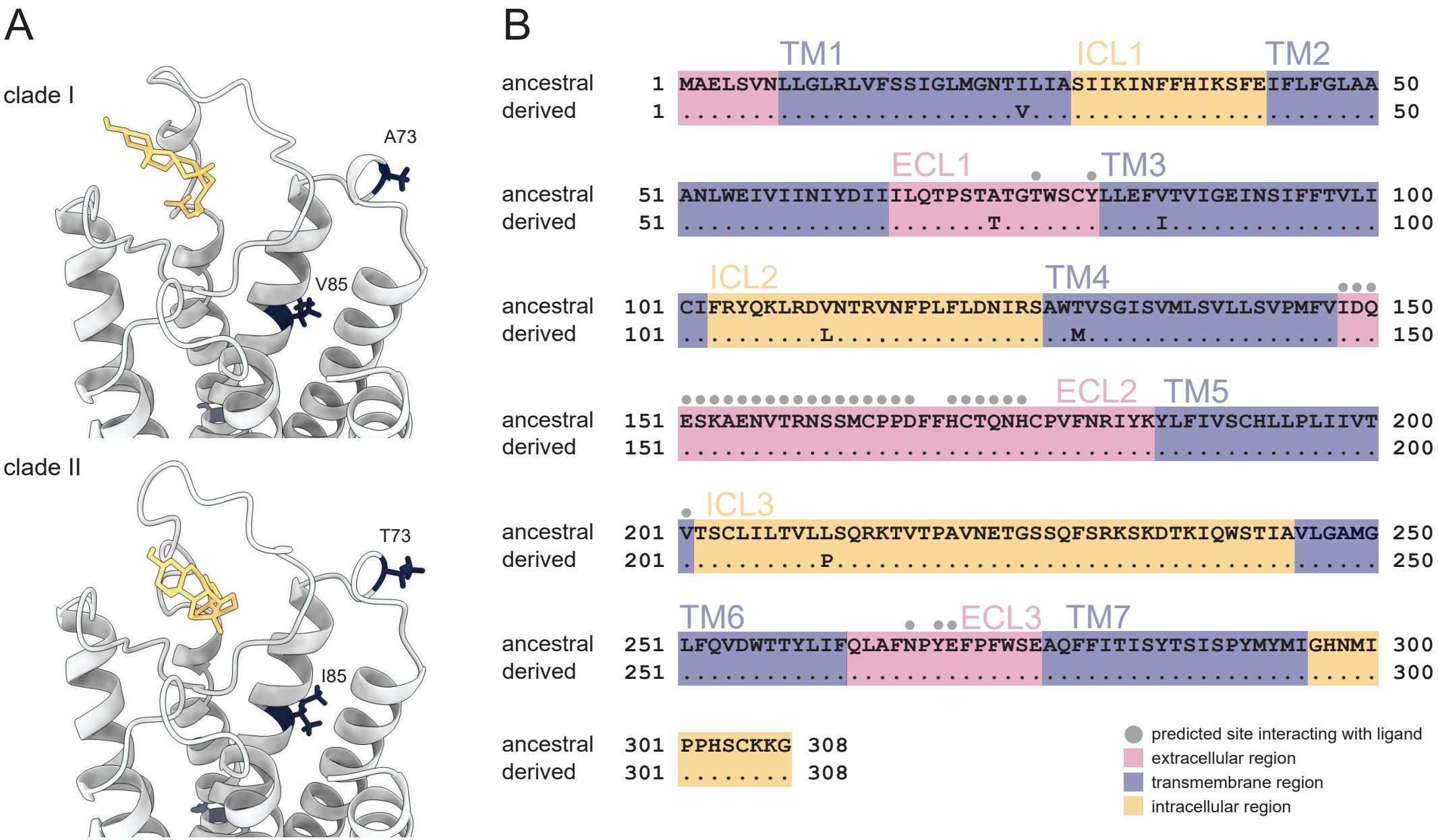

A
clade I
A73
V85
clade II
T73
I85
B
TM1
ICL1
TM2
ancestral 1 MAELSVNLLGLRLVFSSIGLMGNTILIASIIKINFFHIKSFEIFLFGLAA 50
derived 1 .....V..... 50
ECL1
TM3
ancestral 51 ANLWEIVIINIYDIIILQTPSTATGTWSCYLLEFVTVIGEINSIFFTVLI 100
derived 51 .....T.....I..... 100
ICL2
TM4
ancestral 101 CIFRYQKLRDVNTRVNFPLFLDNIRSAWTVSGISVMLSVLLSVPMFVIDQ 150
derived 101 .....L.....M..... 150
ECL2
TM5
ancestral 151 ESKAENVTRNSSMCPPDFFHCTQNHCPVFNRIYKYLFIVSCHLLPLIIVT 200
derived 151 200
ICL3
ancestral 201 VTSCLILTVLLSQRKTVTPAVNETGSSQFSRKSKDTKIQWSTIAVLGAMG 250
derived 201 .....P..... 250
TM6
ECL3
TM7
ancestral 251 LFQVDWTTYLIFQLAFNPYEFPFWSEAQFFITISYTSISPYMYMIGHNMI 300
derived 251 300
ancestral 301 PPHSCKKG 308
derived 301 308
predicted site interacting with ligand
extracellular region
transmembrane region
intracellular region

Figure S15

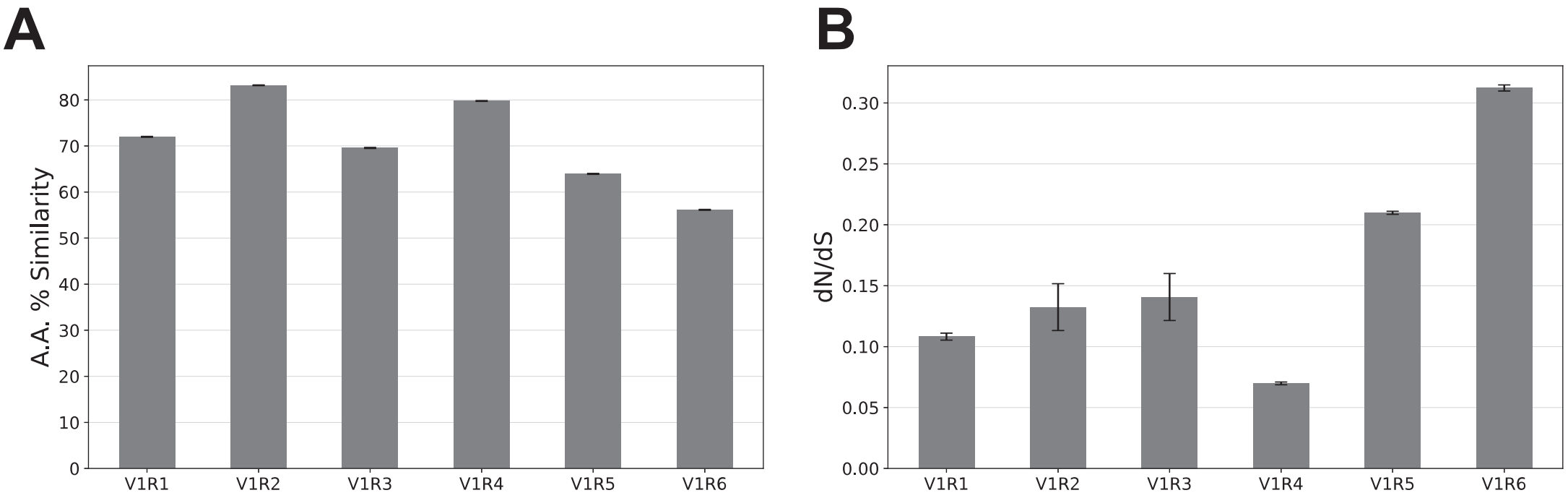

A
A.A. % Similarity
0
10
20
30
40
50
60
70
80
V1R1
V1R2
V1R3
V1R4
V1R5
V1R6
B
dN/dS
0.00
0.05
0.10
0.15
0.20
0.25
0.30
V1R1
V1R2
V1R3
V1R4
V1R5
V1R6

Figure S16