BMC Evolutionary Biology

RESEARCH ARTICLE Open Access

Patterns of genomic differentiation between two Lake Victoria cichlid species, *Haplochromis pyrrhocephalus* and *H.* sp. 'macula'

Shohei Takuno[1†], Ryutaro Miyagi[2,3], Jun-ichi Onami[4], Shiho Takahashi-Kariyazono[1], Akie Sato[5], Herbert Tichy[6], Masato Nikaido[7], Mitsuto Aibara[2], Shinji Mizoiri[2], Hillary D. J. Mrosso[8], Semvua I. Mzighani[2,8], Norihiro Okada[2,9,10*] and Yohey Terai[1,2*†]

Abstract

Background: The molecular basis of the incipient stage of speciation is still poorly understood. Cichlid fish species in Lake Victoria are a prime example of recent speciation events and a suitable system to study the adaptation and reproductive isolation of species.

Results: Here, we report the pattern of genomic differentiation between two Lake Victoria cichlid species collected in sympatry, *Haplochromis pyrrhocephalus* and *H.* sp. 'macula,' based on the pooled genome sequences of 20 individuals of each species. Despite their ecological differences, population genomics analyses demonstrate that the two species are very close to a single panmictic population due to extensive gene flow. However, we identified 21 highly differentiated short genomic regions with fixed nucleotide differences. At least 15 of these regions contained genes with predicted roles in adaptation and reproductive isolation, such as visual adaptation, circadian clock, developmental processes, adaptation to hypoxia, and sexual selection. The nonsynonymous fixed differences in one of these genes, *LWS*, were reported as substitutions causing shift in absorption spectra of LWS pigments. Fixed differences were found in the promoter regions of four other differentially expressed genes, indicating that these substitutions may alter gene expression levels.

Conclusions: These diverged short genomic regions may have contributed to the differentiation of two ecologically different species. Moreover, the origins of adaptive variants within the differentiated regions predate the geological formation of Lake Victoria; thus Lake Victoria cichlid species diversified via selection on standing genetic variation.

Keywords: Cichlids, Population genomics, Adaptation, Speciation, Genomic islands of speciation

* Correspondence: okadanorihiro@gmail.com; terai_yohei@soken.ac.jp
†Shohei Takuno and Yohey Terai contributed equally to this work.
[2]Graduate School of Bioscience and Biotechnology, Tokyo Institute of Technology, 4259 Nagatsuta-cho, Midori-ku, Yokohama 226-8501, Japan
[1]Department of Evolutionary Studies of Biosystems, SOKENDAI (The Graduate University for Advanced Studies), Shonan Village, Hayama, Kanagawa 240-0193, Japan
Full list of author information is available at the end of the article

Background

The molecular basis of the incipient stage of speciation is of great interest in genetics, ecology, and evolutionary biology [1]. Cichlid species in Lake Victoria are a suitable model system to study this stage of speciation [2]. Lake Victoria harbors more than 500 endemic cichlid species [3, 4]. They are thought to have experienced an explosive adaptive radiation during a very short evolutionary period because Lake Victoria dried up at the end of the Pleistocene and was refilled only 15,000 years ago [5, 6]. Indeed, levels of genetic differentiation among species are low, and the species share a large number of nucleotide polymorphisms including differentiated variants shown by studies using a small number of genetic markers, restriction site associated DNA (RAD) data, and whole genome sequencing data [7–14]. Nevertheless, fixed genetic differences between species are expected at loci responsible for adaptive traits and, as a consequence, for speciation. One of the best examples is the long wavelength-sensitive opsin gene (*LWS*), which exhibits a high level of genetic differentiation with fixed genetic differences among Lake Victoria cichlid species [15–19]. As expected, *LWS* alleles are variable among species adapted to different light environments created by different turbidities and different depths [19] and are responsible for speciation by sensory drive [15, 17]. Such variation of *LWS* alleles among species would have originated from the admixture of two divergent lineages [20]. The other gene with fixed genetic differences is the rod opsin gene (*RH1*) for scotopic vision. *RH1* alleles are differentiated among species from different turbidities and depths, and adapted to their ambient light environments [19].

The joint effect of gene flow and divergent selection shapes the pattern of genomic differentiation between an incipient species pair. At the very beginning of this stage, a small part of genes could be involved in reproductive isolation and/or local adaptation [21]. In the latter case, divergent selection acts on these genes, where one allele is advantageous in a species and the other allele is advantageous in its counterpart because separated populations adapt to different niches/environments. Gene flow actually occurs around the target sites of divergent selection but offspring of migrants with the non-adaptive allele are immediately selected out from the species, and as a consequence, the effective migration rate is decreased. On the other hand, gene flow is allowed in other genomic regions and suppresses differentiation, leading to the heterogeneity of genetic differentiation [21–24].

Indeed, the lines of empirical evidence of speciation with gene flow have been recently increased, in plants [25], insects [26, 27], and cichlid species [7, 8, 28, 29]. Despite gene flow, highly differentiated genomic regions between species exist in the genome, and these regions bear genes related to local adaptation such as pigmentation and visual perception in crows [30], beak shape in Darwin's finches [31], and *RH1* for scotopic vision in cichlids [28]. Also, fixed nucleotide differences have been observed in such short genomic regions that emerge only when divergent selection effectively acts [28, 30–34].

Recently, this pattern of genomic differentiation has been reported between two Lake Victoria species, *Pundamillia nyererei* and *P. pundamillia* that both live in rocky habitats [7, 8]. On the other hand, many Lake Victoria cichlid species are distributed along the bottom where there is a soft sandy–muddy substrate, and the pattern of genomic differentiation between species from a sandy–muddy bottom has not been analyzed. The differentiated genomic regions between closely related species from a sandy–muddy bottom may contain candidate genes related to adaptation to micro-habitats, and these candidate genes provide an opportunity to study the incipient evolutionary process in a soft-bottom, benthic ecosystem.

In this study, we focused on two cichlid species from Mwaburugu to reveal the pattern of genomic differentiation between closely related species living in a sandy–muddy habitat. Mwaburugu is a consistently shallow area (2–3 m) with a sandy–muddy bottom, located in the eastern region of Speke Gulf (Fig. 1a). The two species exhibit morphological and behavioral differences (Fig. 1a). *H. pyrrhocephalus* inhabits the middle layer (mainly 7–13 m in Mwanza gulf, south part of Lake Victoria) [35], and *H.* sp. 'macula' is a demersal species [36, 37]. In Mwaburugu, however, the two species distribute in sympatry at a 1- to 3-m depth (Fig. 1a). *H.* sp. 'macula' is a phytoplankton eater, while *H. pyrrhocephalus* is a zooplanktivore [38]. Males of these two species exhibit different nuptial colorations [3], but the distribution of the hue index values largely overlap and are different from other species in Mwaburugu [18]. In cichlids, color perception is important for mate choice [17, 39–41]. Therefore, we expected that the reproductive isolation of these two species may be incomplete, and as a consequence, the genomic differentiation between them may be low. Indeed, we found that most of their genome did not show significant differentiation between the two species.

Results

Summary of population genetic statistics

We performed population genomic analyses in *H. pyrrhocephalus* and *H.* sp. 'macula' from the same locality in Lake Victoria (Fig. 1a). We sampled 20 individuals (10 males and 10 females) for each species, extracted DNA, and constructed Pool-seq libraries (Methods). After mapping the paired-end short reads to the

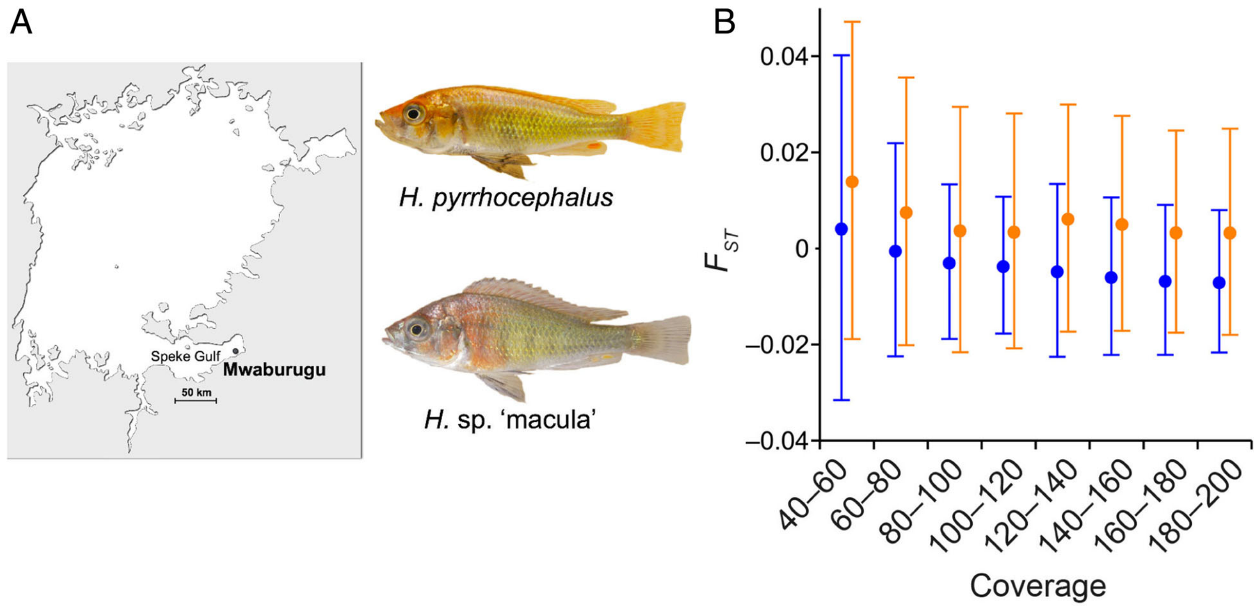

Fig. 1 (**a**) Adult males of *Haplochromis pyrrhocephalus* and *H.* sp. 'macula' with nuptial coloration. A sampling location from the Mwaburugu region is shown by a dot on the map of Lake Victoria. (**b**) Average F_{ST} values (±1 SD) against coverage for Pool-seq data. The blue and orange dots represent observed and simulated values under panmixia

reference genome sequence, the final coverage was ~ 34× in both species. We extracted 4,967,102 and 5,109,870 sites with 80–200× coverage for *H. pyrrhocephalus* and *H.* sp. 'macula,' respectively (Additional file 2: Text S1). We first inferred the site frequency spectrum (SFS) for each species [42]. We estimated 329,613 (6.64%) and 322,956 (6.32%) polymorphic sites in *H. pyrrhocephalus* and *H.* sp. 'macula,' respectively. The folded SFSs were very similar between the two species (Additional file 1: Figure S1A), and almost 70% of segregating sites were inferred to be singletons in each species. Nucleotide diversity [43] was 0.00717 and 0.00678 for the two species. Tajima's *D* values [44] were highly negative (less than − 2 in both species), indicating an excess of rare alleles in each species. This observation suggests population expansion [45] accordingly, and we inferred demographic parameters based on the SFSs using δa/δi [46] (Additional file 2: Text S1). As expected, we detected rapid population expansion after the colonization of Lake Victoria; the current effective population size was on the order of 10^6 (Additional file 2: Text S1).

Note that the power and accuracy of a Pool-seq analysis would be low for allele frequency estimation, especially for alleles at lower frequency [47]. Nevertheless, nucleotide diversity and Tajima's *D* values are consistent with previous estimates based on the Sanger method [12, 48]. We further note that we utilized sites with a much higher coverage than the genomic average, and high coverage can indicate repetitive regions. After excluding sites within repetitive regions, we obtained nearly the same statistics for each species (Additional file 2: Text S1).

Two Lake Victoria cichlid species would be close to a single panmictic population

We measured the level of population differentiation between the two species. We estimated the allele frequency at every site given the estimated rate of sequence errors (Materials and Methods, Additional file 2: Text S1). We used 7,516,409 polymorphic sites, at which the coverage was ≥40 for both species. When the minor allele frequency in the total dataset was higher than 0.05, > 80% of sites exhibited shared polymorphisms, as observed in Lake Victoria cichlids [7–14, 48], suggesting that the species are closely related. We applied F_{ST}statistics [49], which is suitable to quantitatively assess the level of differentiation between such closely related species or populations. This statistic is calculated by.

$$F_{ST}=1-\frac{2T_W\mu}{2T_B\mu}=1-\frac{\pi_W}{\pi_B} \tag{1}$$

where T_{W} and T_{B} represent the coalescent time of samples from the same population and that of samples from different populations, respectively; μis the mutation rate per generation, and π_{W} and π_{B} represent the average nucleotide diversity within each population and the average pairwise nucleotide divergence between the two populations, respectively. To measure the level of population differentiation, T_W was used as a control. When two sets of samples are from the same population, T_W and T_B are expected to be equal and F_{ST} is close to 0.

We calculated F_{ST} for each of the 7,516,409 segregating sites. Because the accuracy of allele frequency estimates is low when coverage is low [47], F_{ST} is expected to become relatively high at such sites. As expected, we found a slight, but detectable, negative correlation between F_{ST} and coverage (blue plot in Fig. 1b). Nevertheless, average F_{ST} values were around zero even in alleles with low coverage (< 0.0043; Fig. 1b), indicating almost no population differentiation.

We examined whether we could treat the two species as a single panmictic population. In theory, when the migration rate is high enough (i.e., when the population migration rate, $4Nm$, is higher than 10, where N is the effective population size and m is the migration rate per gamete per generation), the pattern of nucleotide polymorphisms in samples from two populations is expected to be similar to that in samples from a single population [50]. To test this, a standard coalescent simulation is not sufficient because we used Pool-seq data. Thus, to generate null distributions of F_{ST} values given our coverage, we simulated both the coalescent process under panmixia and the Pool-seq process with the inferred population expansion (Additional file 2: Text S1). The distributions of simulated F_{ST} values exhibited a similar tendency (orange plot in Fig. 1b), though the observed distributions were slightly skewed toward negative values. When a single segregating site was analyzed, F_{ST} statistics became negative when allele frequencies in the two sample sets were very similar. This indicates a slight excess of the proportion of segregating sites with very similar allele frequencies between the two species. The cause of this excess is not known, but the observed and expected distributions were largely overlapping. Thus, we reasoned that the two species might be close to a single panmictic population (see also Fig. 2a).

We have three caveats in this section. First, as in the previous section, we excluded sites in repetitive regions, and we obtained essentially the same results as in Fig. 1b (Additional file 2: Text S1). The second caveat is that a Pool-seq analysis is less powerful for detecting population differentiation, especially when minor allele frequency is low [47]. However, in our case we could detect population differentiation when F_{ST} was ≥0.01, and when the sample size was set to be 40 in each population (Additional file 2: Text S1). Furthermore, we excluded singletons and obtained essentially the same results as in Fig. 1b. Finally, the females of *H.* sp. 'macula' are morphologically similar to other *Haplochromis* or *Enterochromis* species, and it is possible that these female specimens included misidentification of species. To avoid the misclassification of species, we performed the same analysis only in male individuals. Despite the reduction in sample size, we obtained essentially same result (Additional file 3: Figure S2).

Identification of highly differentiated genomic regions

Despite minimal genomic differentiation, *H. pyrrhocephalus* and *H.* sp. 'macula' exhibit differences in morphology and habitat [3]. Thus, we performed a genome scan to search for candidate genes that are subject to divergent selection using sites with 20–200× coverage in both species (Additional file 2: Text S1). We used 10-kb window size with 2-kb increments for the scan, and calculated π_W, π_B, and F_{ST} for each window. The spatial distributions of π_W and π_B are almost the same across the genome, indicating that T_W and T_B are almost the same and F_{ST} values are close to zero (see two typical genomic regions in Fig. 2a) due to the panmixia of the two species (Fig. 1b). On the other hand, the *LWS* gene, a prime example of target genes under divergent selection [15, 17, 18], exhibited a remarkable pattern. The spatial patterns of π_W, π_B, and F_{ST} around *LWS* are shown in Fig. 2b. We observed that π_B values are significantly higher than π_W, and as a consequence, there is a clear peak of F_{ST} in the 14-kb region that includes the *LWS* gene. Despite the low accuracy of Pool-seq, these statistics are very close to the estimates reported in a previous study, which determined the sequences by the Sanger method [18]. Outside the 14-kb region, π_W and π_B values are almost equal as in Fig. 2a. Furthermore, we found seven fixed nucleotide differences between *H. pyrrhocephalus* and *H.* sp. 'macula' within the peak (green triangles in Fig. 2b) that indicate strong signatures of divergent selection under the pressure of extensive migration.

To screen candidate genes under divergent selection, we initially filtered windows with the top 0.1% of F_{ST} values ($F_{ST} > 0.372$). We performed a neutrality test to see if such high F_{ST} values are observed without natural selection. We simulated a coalescent process and Pool-seq process with 20× coverage to maximize the variance of a null distribution (Additional file 2: Text S1), and obtained a very small P-value ($P < 10^{-5}$; false discovery rate < 0.014). We further screened such windows that exhibited a clear peak of F_{ST} values as in *LWS* genes (Fig. 2b) and fixed nucleotide differences within the peaks. In total, we detected 21 highly differentiated regions (14–28 kb). Hereafter, we focused on these 21 short differentiated regions (DRs), and three examples of DRs are shown in Fig. 2c-e and the others in Additional file 4: Figure S3. We searched for genes in the DRs by BLASTN and explored their biological roles in terms of adaptation and speciation. Nineteen out of 21 DRs included 1–3 genes (Table 1), and 28 total genes were found in the DRs.

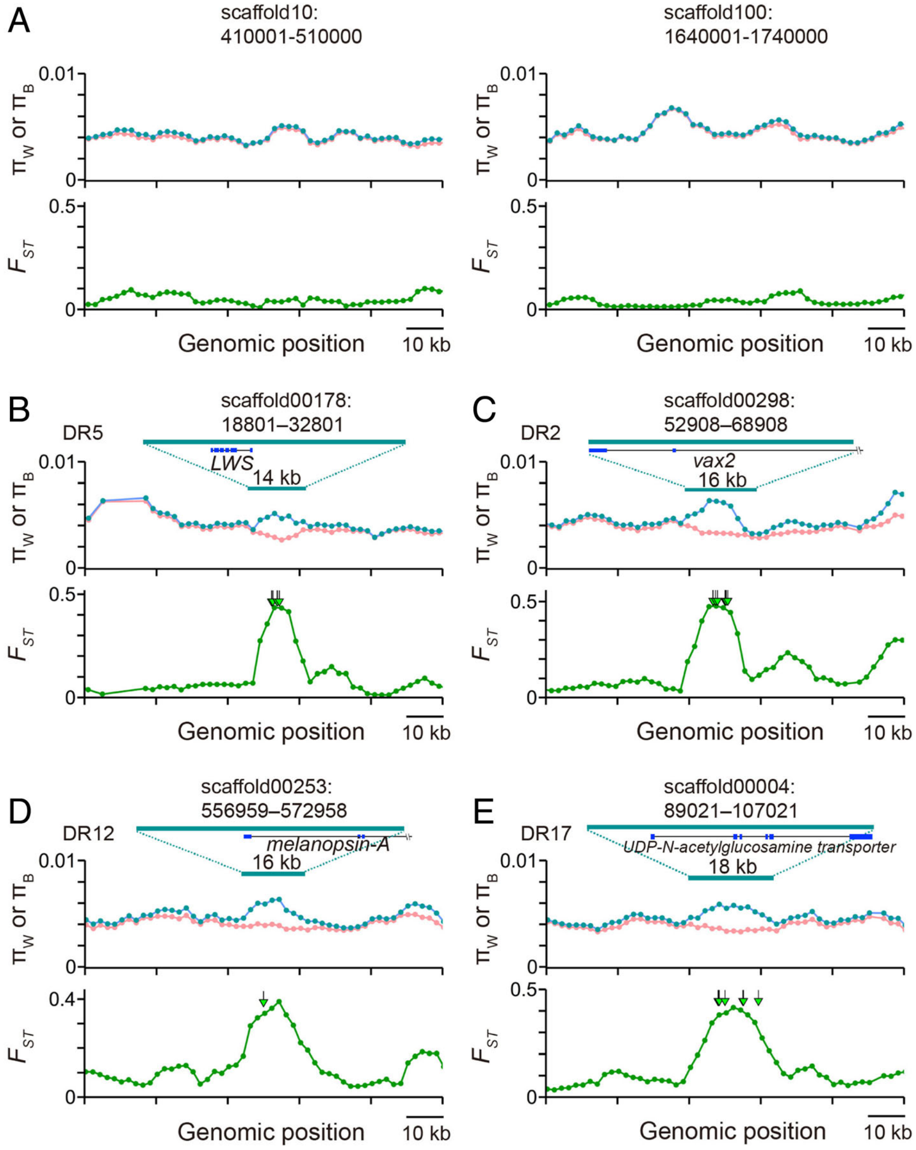

Fig. 2 The spatial patterns of the average values of intraspecies nucleotide diversity in *H. pyrrhocephalus* and in *H.* sp. 'macula' (π_W; pink); average pairwise nucleotide divergence between the species (π_B; blue); and F_{ST} (green). (**a**) Typical genomic regions. (**b**–**e**) Candidates of the target genes of divergent selection. The green arrows represent fixed nucleotide differences between species. Green and blue solid lines indicate differentiated regions (DRs) and exons of genes, respectively. The positions of DRs are described on the right side of the green lines

As mentioned above, allele frequency estimates in regions of low coverage are not accurate, especially for alleles with low frequencies [47]. Thus, we repeated the analysis, discarding segregating sites with minor allele frequency ≤ 0.05, and identified the same set of 21 regions. Furthermore, we verified the fixed nucleotide differences by Sanger sequencing for 5 of 21 loci (we selected a subset of 5 DRs due to limited amounts of DNA samples) (Additional file 1: Figure S1C). To avoid the misclassification of species, we repeated the analysis using only male samples as in the previous section. We obtained essentially the same result (Additional file 3: Figure S2B) as in the *LWS* genes (Fig. 2b) and in other DRs.

The pattern of polymorphisms around the target site of divergent selection

The hitchhiking effect of divergent selection under the pressure of migration is more limited than that of positive selection (i.e., selective sweep). To show this, we performed a population genetic simulation under a simplified model (Methods). We assumed an isolation with migration model with two populations (populations 1 and 2), into which we incorporated mutation,

Table 1 Genes in DRs

DRs	Gene name	Predicted functions related to adaptation and speciation [c]
DR1[a]	diaphanous	Developmental process, cell movement, auditory
DR2[a]	ventral anterior homeobox 2	Development of retina
DR3[a]	prostaglandin d2 receptor 2	Unknown
	G-protein coupled receptor 4 (GPR4)	Adaptive to different oxygen concentration
	UDP-glucuronosyltransferase 2b15	Unknown
DR4	hemicentin-1	Unknown
DR5[a]	long wavelength-sensitive opsin (LWS)	Speciation by sensory drive
DR6[a]	netrin receptor UNC5c	Brain development
DR7	general transcription factor IIH subunit 1	Unknown
DR8[a]	intestinal mucin	Host-specific microbiota composition
DR9[a]	hepatocyte growth factor receptor	Morphogenesis for the muscles of fins, affecting mobility
DR10[a]	*tbx3*[b] (30 kb downstream from DR10)	Developmental process
DR11[a]	ap-4 complex subunit epsilon	Unknown
	cytochrome p450 aromatase type II	Sexual differentiation of the brain and reproductive behavior
	gliomedin	Unknown
DR12[a]	melanopsin A	Photic regulation of circadian clocks
DR13	No gene	–
DR14	Uncharacterized protein	–
	Uncharacterized protein	–
DR15	Uncharacterized ncRNA	–
DR16[a]	aryl hydrocarbon receptor nuclear translocator	Adaptive to different oxygen concentrations
DR17	UDP-n-acetylglucosamine transporter	Unknown
	U3 small nucleolar ribonucleoprotein protein imp3	Unknown
DR18[a]	peptidyl-prolyl cis-trans isomerase H	Unknown
	transcription initiation factor TFIID subunit 10	Early embryonic development
	G-protein coupled receptor 160	Unknown
DR19[a]	type II cytoskeletal 5	Epidermis development
DR20[a]	hydroperoxide isomerase aloxe3	Epidermis development
	macrophage mannose receptor 1	Unknown
DR21[a]	Ras-related protein rab-11a	Unknown
	RNA-binding protein mex3a	Brain aging

[a] DRs contained genes with predicted roles for adaptation and speciation
[b] This DR did not contain a gene, but *tbx3* was located 30 kbp downstream from this DR
[c] References are listed in Additional file 6: Table S1

recombination, migration, divergent selection, and random genetic drift. The locus I is subject to divergent selection, and has two alleles, A and a. The A allele is advantageous in population 1, and the a allele is advantageous in population 2. At the start of simulations, locus I is monomorphic for the A allele, and a linked neutral locus (locus II) is under mutation-drift equilibrium. At the time $T = 0$, we introduced the a allele in population 2 with the initial frequency of $1/2N$ and ran simulations for $8N$ generations. The beneficial alleles (allele A in population 1 and allele a in population 2) are expected to immediately reach an equilibrium frequency, $\hat{p}$, when population size is infinite:

$$\hat{p} = -\frac{m}{s} + \frac{1}{2} + \sqrt{\left(\frac{m}{s}\right)^2 + \frac{1}{4}} \tag{2}$$

where m is migration rate per generation, and s is the selection coefficient. As long as s is much higher than m (e.g., $s/m >> 5$), $\hat{p}$ is very close to 1. We ran simulations 100,000 times each with a variety of pairs of m and s, and calculated expected π_W, π_B, and F_{ST} at several time

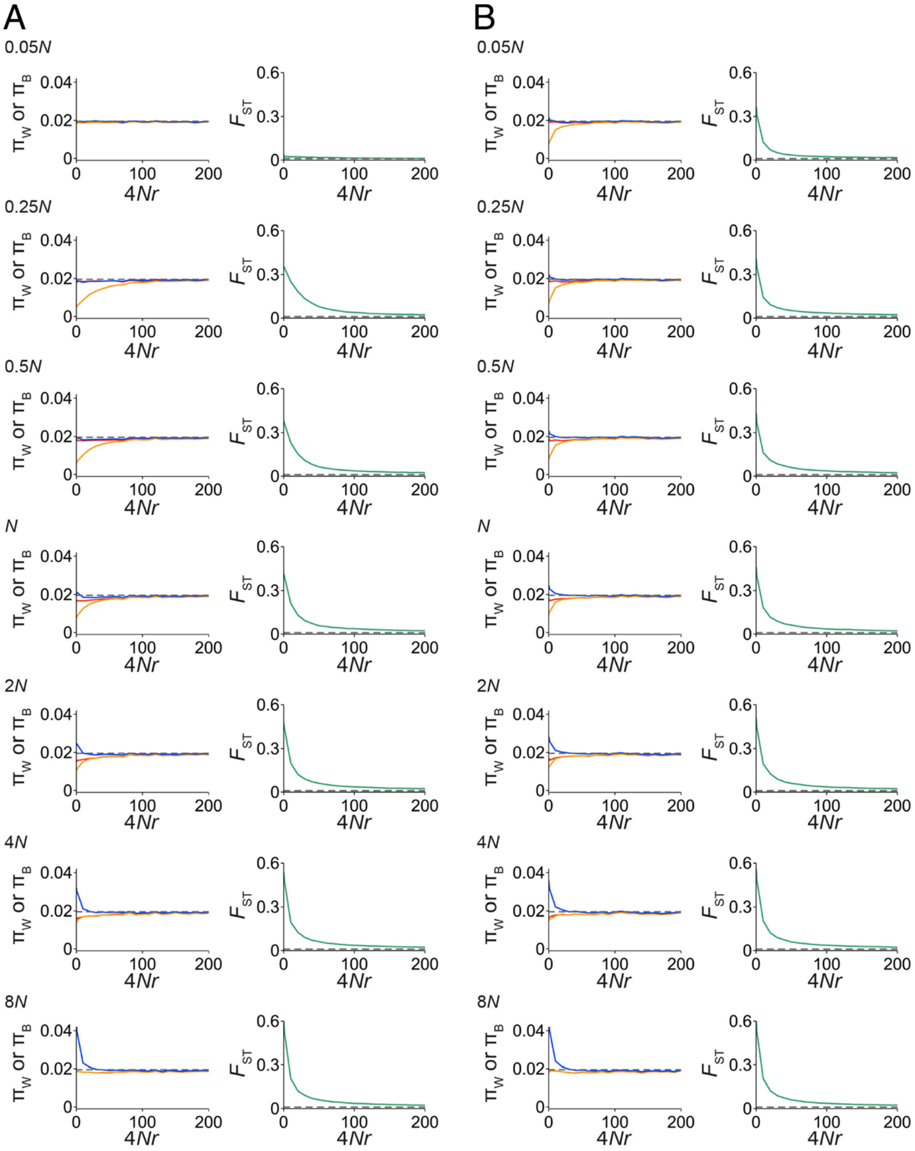

Fig. 3 The hitchhiking effect of divergent selection when a new beneficial allele arises by a new mutation (**a**) and is derived from standing variation (**b**). The *x*-axes represent distance from the target site of divergent selection (scaled by the population recombination rate), and the *y*-axes represent nucleotide diversity in population 1 (π_W; red), π_W in population 2 (orange), average pairwise nucleotide divergence between species (π_B; blue), and F_{ST} (green). Dashed gray lines indicate the theoretical expectations of π_W and F_{ST} under neutrality

points. We confirmed that the frequencies of the beneficial alleles in both populations quickly reached $\hat{p}$ in finite populations, and the patterns of polymorphisms were qualitatively consistent among the pairs of *m* and *s*.

We show the result with $4Nm = 50$, and $4Ns = 400$ in greater detail in Fig. 3a. First, the *a* allele quickly reaches $\hat{p}$ in population 2 which causes a strong reduction of π_W in long genomic regions such as in the case of a selective sweep event [51, 52]. However, the signature of selective sweep is quickly eliminated by the effects of migration and recombination, and a short differentiated region appears with fixed nucleotide differences as shown in Fig. 2b-e and Additional file 3: Figure S3. This situation is very similar to the process of neofunctionalization of duplicated genes (an analog to divergent selection in our model) under the pressure of interlocus gene conversion (analogs to migration and recombination), where a short differentiated peak between duplicates appears around the target site of neofunctionalization [53]. We further simulated the situation in which the *a* allele is derived from standing genetic variation because it has been proposed that adaptive variants are derived from standing variation ([20]; see the section “Lake Victoria cichlid species diversified via selection on standing genetic variation" below). We ran simulations with $s = 0$ until the frequency of the *a* allele reached 20%, and then divergent selection started to act. The result is shown in Fig. 3b, and we found that the shrinkage of the differentiated region is faster than that in Fig. 3a.

A high false positive rate for the F_{ST}-outlier approach has been pointed out previously, and using absolute nucleotide divergence (or π_B) is recommended [54, 55]. F_{ST} can be increased without divergent selection when π_W is reduced as in Eq. (1), for example, by a classical selective sweep or by background selection. When one focuses on highly differentiated species, we may not be able to ignore this flaw of the F_{ST}-outlier approach (i.e., when T_B is much older than T_W; $F_{ST} > 0.1$ as in [54]). However, a high false positive rate is expected when using π_B in low-differentiated, young species pairs with extensive migration such as *Haplochromis* species (Figs. 1b and 2). This is because high π_B regions occasionally appear due to the large variance of coalescent time in the ancestral population. If such regions have evolved in a neutral manner, π_W values are expected to be equal to π_B due to migration, and F_{ST} is close to zero. Thus, employing F_{ST} should perform better in our case.

In our DRs, the possibility of a classical selective sweep without divergent selection is unlikely. As shown in Fig. 3, we observe the sweep-like strong reduction of π_W that actually increases F_{ST}. However, without divergent selection, such reduction of π_W is quickly eliminated. Both divergent selection and migration are required to maintain short differentiated regions as observed in Fig. 2b-e and Additional file 3: Figure S3. Background selection is also unlikely. The effect of background selection is remarkable on regions with low recombination rates. Purifying selection purges deleterious mutations, and also linked neutral variants together. As a consequence, π_W is slightly reduced in relatively long genomic regions [56], and F_{ST} values become high. However, the lengths of DRs are very short (14–28 kb; Fig. 2b–e; Additional file 3: Figure S3). Furthermore, the inferred population sizes in *Haplochromis* species are fairly large (Additional file 2: Text S1), and therefore the population recombination rate ($4Nr$, where N is effective population size and r is recombination rate per generation) is so high that recombination effectively reduces the effect of background selection. More importantly, in both cases, we would not expect to observe fixed nucleotide differences under the pressure of migration without divergent selection, however, we did (Fig. 2b-e, Additional file 3: Figure S3).

The roles of fixed differences in DRs

Among the genes in DRs (Table 1), six nonsynonymous fixed differences were located in the coding region of *LWS* (Additional file 1: Figure S1C). These nonsynonymous substitutions make 8 and 9 nm shifts in the absorption spectra of LWS photo-pigments, with 11-*cis* retinal (A1-) and 11-*cis* 3- dehydroretinal (A2-derived retinal), respectively [18], suggesting that the fixed differences are responsible for the functional difference of LWS pigments between two species.

In contrast, four fixed differences in another opsin, *melanopsin A*, were not located in the coding region, but are in the upstream region of the gene (Fig. 4a), raising the possibility of differential expression. We quantified the expression level of this gene in the eyes of six individuals each from *H. pyrrhocephalus* and *H.* sp. 'macula' by real-time qPCR. We detected a significant difference in expression between species (Fig. 4a), indicating the possibility that these substitutions may cause the expression difference of *melanopsin A* in these two species.

The color pattern of cichlids is one of the most variable traits among species in Lake Victoria; therefore, we performed a pooled RNA-seq analysis to screen for the candidates of differentially expressed genes in the anterior part of the lateral skin of these two species. In total, 172 contigs showed differential expression between the two species ($P < 0.05$). These included three genes in DRs, *P450 aromatase*, *UDP-N-acetylglucosamine transporter*, and *aloxe3*. The fixed differences in these genes were also located in their upstream regions (Fig. 4b–d). Therefore, we quantified the expression levels of these genes in the anterior part of the lateral skin of ten individuals each from both species. In all three genes, the expression levels in the anterior part of the lateral skin were completely different between the two species (Fig. 4b–d, $P < 0.001$). These results also indicate the possibility of differential expression of genes caused by the fixed differences.

Lake Victoria cichlid species diversified via selection on standing genetic variation

We inferred the origin of the putative adaptive variants (i.e., fixed differences) in cichlid species. We determined the sequences (~ 1 kbp), including the fixed differences, within 16 DRs from Lakes Victoria, Malawi, Tanganyika, and riverine *haprochromis* species. We constructed phylogenetic trees based on our sequences and orthologous sequences in other cichlid genomes [14]. No tree for any DR showed monophyly of Lake Victoria species (Fig. 5a and Additional file 5: Figure S4A), suggesting that these alleles in DRs arose as presumably neutral variants in the ancestral population of the two species. By contrast, five trees showed monophyly of the Lake Victoria superflock, including species from Lake Victoria and surrounding rivers [9, 16] (Additional file 5: Figure S4F), suggesting that the adaptive variants arose in the common ancestor of this monophyletic clade (riverine origin: Fig. 5b and Additional file 5: Figure S4B). The

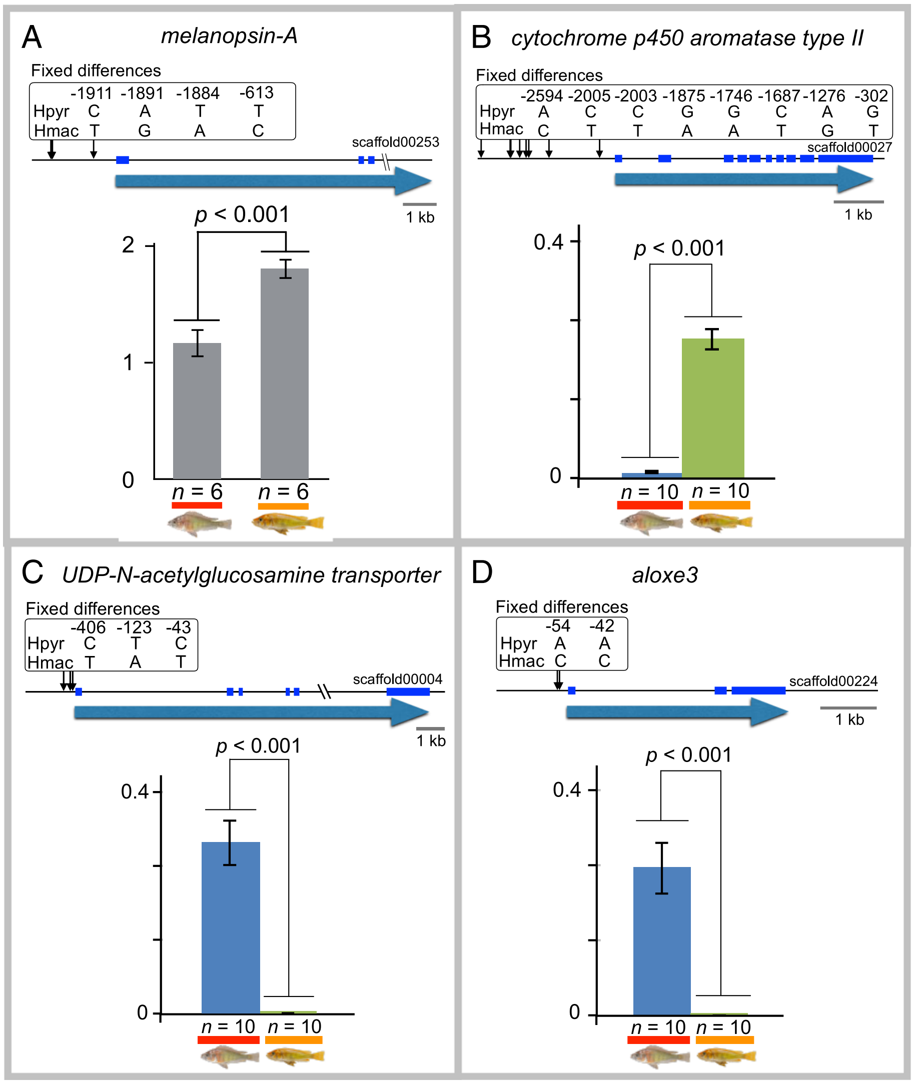

Fig. 4 (**a**–**d**) Different expression levels of genes between *H. pyrrhocephalus* (Hpyr) and *H.* sp. 'macula' (Hmac). Expression was analyzed by qPCR using RNA from eyes (**a**, each of three individuals) and the anterior part of lateral skin (**b**–**d**, 10 individuals each). Gene names are shown at the top of the panels. The direction of blue arrows indicates the gene direction. Black arrows and blue solid lines represent the positions of fixed differences and exons of genes, respectively. The numbers and nucleotides under the numbers in the rectangles indicate the position from the initiation codon and fixed nucleotides in *H. pyrrhocephalus* (upper) and *H.* sp. 'macula' (lower), respectively

origin of variants in these five DRs are coincident with the scenario that two divergent lineages admixed in the ancestor of the Lake Victoria superflock [20]. *Astatotilapia burtoni* and species from Lake Victoria, surrounding rivers, Lake Malawi and tribe Tropheini (Lake Tanganyika) were mixed in a monophyletic clade in the other 11 trees (Additional file 4: Figure S4E), suggesting that the mutations accumulated in the common ancestor of this lineage [Fig. 5c and Additional file 4: Figure S4C, modern haplochromines origin [57]]. Hence, the origin of the adaptive variants can be traced back to some point in the ancestral lineage of modern haplochromines.

Discussion

In this study, we focused on *H. pyrrhocephalus* and *H.* sp. 'macula,' which are distributed in sympatry in Mwaburugu (Fig. 1a) where the bottom is a sandy–muddy substrate with no rocks, woody debris, or other structures.

In the genome scan, we detected 21 DRs (14–28 kb) including 28 genes. *LWS*, located in DR5 (Fig. 2b), is involved in adaptation to different light environments [19] and contributes to speciation by sensory drive [15, 17], suggesting that DRs contain genes responsible for cichlid adaptation and reproductive isolation. DR2 and DR12 contain the vision-related genes ventral anterior

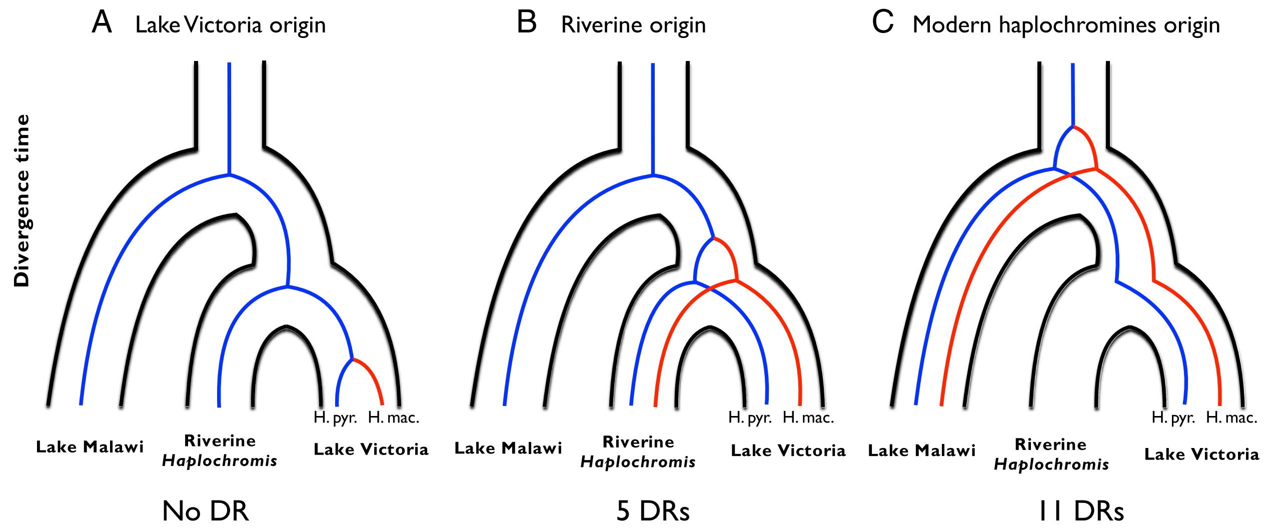

Fig. 5 The origins of mutations in DRs. Three genealogies represent the accumulation of mutations tracing back to the common ancestral species of (**a**) Lake Victoria species, (**b**) Lake Victoria and riverine *Haplochromis* species, and (**c**) Modern haplochromines. Blue and red lines indicate *H. pyrrhocephalus* and *H.* sp. 'macula' alleles, respectively

homeobox 2 (*vax2*) and melanopsin A, respectively (Fig. 2c and d; Table 1). The *vax2* gene is involved in the regulation of retinal development [58] and melanopsin A is involved in the photic regulation of the circadian clock [59]. *H. pyrrhocephalus* and other pelagic zooplanktivorous species (e.g. *H. laparogramma, H. heusinkveldi*) migrate toward the surface in the evening and stay during the night to forage for zooplankton [35]. *H. pyrrhocephalus* has very large double cone photoreceptor cells in their retinas for high sensitivity to light [60]. The genes involved in retinal development and circadian clock regulation may contribute to the specific features of *H. pyrrhocephalus.* Although we did not analyze other pelagic zooplanktivorous species, DR2 and DR12 might be shared among ecologically similar species with *H. pyrrhocephalus.* Other DRs contained several genes involved in developmental processes, such as the development of the brain (DR6), epidermis (DRs 1, 19, and 20), and fin muscles (DR9) (Table 1, references in Additional file 6: Table S1). The importance of brain activity for sociality and reproduction was reported in cichlids [61] . Both the epidermis structure and fin muscles may be related to species-specific characteristics; the epidermis directly interacts with the external environment and fin muscles affect fish mobility. P450 aromatase (DR11) is responsible for estrogen synthesis and plays a regulatory role in sex determination, gametogenesis, central nervous system development, and reproductive behavior [62], which are important traits for sexual selection. Interestingly, Atlantic cod populations are also differentiated at this gene [63]. Chronic hypoxia has been observed in Lake Victoria [64]. Cichlid species adapt to hypoxia by multiple strategies [65]. Aryl hydrocarbon receptor nuclear translocator gene (DR16) is involved in physiological adaptation to hypoxia [64] and the G-protein coupled receptor 4 (DR3) regulates breathing by CO_2 stimulation [65]. Lake Victoria cichlid species may experience hypoxia in deep water, heavily vegetated shallow shoreline habitats, and dense algal blooming in open water [66]. Since *H. pyrrhocephalus,* except in Mwaburugu, inhabits deeper water (inferred to be hypoxic) than *H.* sp. 'macula,' these genes may be involved in adaptation to different oxygen concentrations. Intestinal mucin (DR8) functions as a host-specific determinant affecting the gut microbiota composition, which is important for digestion [67]. Intestinal mucin may be involved in the digestion of species-specific food such as phytoplankton and zooplankton in *H.* sp. 'macula' and *H. pyrrhocephalus*, respectively. *Mex3a* (DR21) is associated with brain aging in the short-lived fish *Nothobranchius furzeri*, a model of aging studies [68], and may be related to interspecific differences in aging in Lake Victoria cichlids. In total, we detected genes with predicted roles in adaptation and speciation within at least 15 out of 21 DRs (Table 1), suggesting that the DRs contain genes responsible for adaptation and reproductive isolation. Two DRs (DR10 and DR13) did not contain genes, but might contain regulatory regions of genes (e.g., DR10 was located 30 kb upstream of *tbx3*).

How are the observed fixed differences related to functional differences in genes in DRs? The nonsynonymous fixed differences in *LWS* cause 7-nm shifts in the absorption spectra of LWS pigments [18]. The fixed differences in four genes (melanopsin A, P450 aromatase, UDP-*N*-acetylglucosamine transporter, and *aloxe3*) were located in the upstream regions of the genes and are associated with significant differences in expression between the two species. P450 aromatase expression in the anterior part of lateral skin might partly explain the difference in regulation of color pattern formation between the two species (Fig. 1a); this gene plays a regulatory role in various sexual traits [62]. The different expression levels of these four genes suggest that the fixed differences affect gene expression levels (Fig. 4). Hence, the fixed differences observed in DRs are expected to affect protein functions or gene expression. These results support the hypothesis that the divergence of Lake Victoria cichlid species is explained by differentiation in short genomic regions containing genes responsible for adaptation. Each of the short genomic regions may be responsible for the adaptive traits, and the combination of these traits including LWS adaptation may lead to reproductive isolation between ecologically different species.

It has been argued that standing genetic variation is important for the recent radiation of cichlid lineages based on genome-wide shared polymorphisms among different cichlid lineages [13, 14, 69], many of which are likely selectively neutral. Recently, Meier et al. reported that the ancient admixture event between distantly related lineages (Congolese and Upper Nile) would increase genetic variation that would have contributed to the morphological diversity and adaptive traits in the Lake Victoria region superflock [20], suggesting the importance of standing genetic variation. However, the origins of adaptive variants are not fully elucidated. We examined whether or not the origin of adaptive variants in our DRs are the same age as the two divergent lineages in Meier et al. 2017 by tracing back to the origins of the putative adaptive variants (i.e., fixed differences in DRs) in cichlid species. The phylogenetic trees based on the five DR sequences were in agreement with this prediction. Furthermore, the origins of the other 11 DRs were older than the common ancestor of the Lake Victoria region superflock. These estimations suggest that the adaptive variants in DRs can be traced back to some point in the ancestral lineage of modern haplochromines [57].

In this study, we demonstrate extensive gene flow between ecologically differentiated species, 21 DRs, and an ancient origin of the adaptive variants responsible for species divergence. These findings provide new insight into a long-standing question: why do cichlids represent the most successful radiation in Lake Victoria during a very short evolutionary period? The ancient origin of adaptive variants within DRs provides an important clue. These mutations tended to accumulate after the split of the modern haplochromine lineage, supporting the idea that the radiation of Lake Victoria species occurred via selection on standing genetic variation [20]. It may also be explained by extensive gene flow, which enables species to share functional sequences that may have promoted adaptation to various environmental conditions within Lake Victoria. Indeed, the same *LWS* allele that is adaptive to local light environments has been observed in multiple species [15, 18, 19]. In this study, we focused on two species in Lake Victoria. Additional genomic sequences of multiple individuals from additional species pairs with extensive gene flow will allow us to paint a more comprehensive picture of the role of gene flow in Lake Victoria cichlid radiation.

Methods

Sample information

Two Lake Victoria cichlid species were used, i.e., *Haplochromis pyrrhocephalus* Witte and Witte-Maas (1987) [70] and *H.* sp. 'macula.' These species are widely distributed in Lake Victoria [3, 71] and inhabit Mwaburugu at the east end of Speke Gulf (Fig. 1a). All specimens were collected in sympatry by netting (1.5-m height) at a 1- to 3-m depth in Mwaburugu. All fish were collected by M.A. and S.M. in 2004–2006. The identification of all specimens was verified by M.A. and S.M.

Species identification: *Haplochromis pyrrhocephalus* is one of the most common pelagic-sublittoral species in the eastern Speke gulf. The species is recognized to be in the *Yssichromis* group because of the slender body (body depth 27.5–31.1% of standard length in the original description and 26.2–30.3% in our measurement, see material and methods in [72]). This species is distinguished from all other Haplochromines by a combination of the slender body and male nuptial coloration: 1) orange to red coloration on the head, unpaired fins, and egg dummies, and 2) absence of a lateral band. We collected 14 species from Mwaburugu (Fig. 1a), southeastern Speke gulf, where no slender-bodied species were found except this species. Therefore, we identified slender-bodied females within the range described above as *H. pyrrhocephalus.*

Haplochromis sp. 'macula' was described by the *Haplochromis* Ecology Survey Team (HEST) in Leiden University and subsequently re-described by Seehausen [3] with male nuptial coloration. Male of the species is relatively easy to identify because of the bright red coloration on the head, anterior body, and dorsal fin membrane, and yellow to green coloration on the posterior body and caudal peduncle. Among all species that we collected in Mwaburugu, *H.* sp. 'macula' is morphologically different from the other species by the combination of the following traits: 1) dorsal head profile is straight or weakly moderately curved (vs. moderately curved in the other species); 2) oral teeth in outer jaw are weakly compressed (vs. cylindrical to weakly compressed); 3) flange of main cusp of oral teeth in outer jaw are relatively prominent (vs. without or weak flange); and 4) arrangement of anterior teeth in outer jaw is relatively dense, with posterior end of the tooth slightly overlapping to the anterior end of the neighboring tooth (vs. not overlapping). In this study, we chose females which possessed all of these characteristics as *H.* sp. 'macula'.

Additional genetic information: all specimens used in the present study were subjected to the species identification procedure described above prior to the analysis of opsin genes [18]. Among all species that we collected in Mwaburugu, one *LWS* gene allele Py and H was exclusively fixed in *H. pyrrhocephalus* and *H.* sp. 'macula', respectively. In particular, the Py allele was only found in *H. pyrrhocephalus*. Thus, these two species possess a genetic biomarker specific to species. Note, however, that we did not identify these species by using genetic information.

Pooled genomic DNA sequencing (Pool-seq) and mapping

Genomic DNAs were extracted from caudal or pectoral fins of wild-caught individuals using the DNeasy Blood & Tissue Kit (Qiagen, Hilden, Germany). All tissues were dissected and kept in 100% ethanol until use. Equal amounts of DNAs (500 ng) extracted from 10 males and 10 females each from *H. pyrrhocephalus* and *H.* sp. 'macula' were pooled. In each species, DNAs from males and females were pooled separately. Libraries were constructed using the TruSeq DNA LT Sample Prep Kit (Illumina, San Diego, CA, USA) and the sequences were determined (paired-end 100 bp) using the Illumina HiSeq2000 platform. The paired-end short reads were mapped to the reference genome sequence of *Pundamilia nyererei* [14] using Bowtie 2 [73] specifying "--score-min L,0,-0.2."

The number of high-quality alleles was counted at every site using Samtools with the arguments "-C50 -q20 -Q30" [ver. 0.1.19-44,428 cd; [74]]. If an indel was called, the site was filtered out, including the regions 9 bp upstream and downstream of the site. Sites with ≥3 nucleotides were removed. Sites with coverage of ≥20× were retained. Finally, if PV4 information was available, a site was filtered out if the *P*-value for strand bias or tail distance bias was less than 10^{-4} or if the *P*-value for baseQ bias was less than 10^{-100} according to the default settings in VCFtools [75].

Population genomic analyses

Pool-seq data for each species were initially analyzed separately. Mono- and bi-allelic sites with coverage of 80–200× were screened for each species to calculate the site frequency spectrum (SFS) by applying the EM algorithm developed by Boitard et al. [42] (Additional file 2: Text S1). The demographic history of each species was inferred from the SFS using δa/δi [46] (Additional file 2: Text S1).

Population differentiation between the two cichlid species was analyzed. Sites with coverage of 20–200× were extracted for both species, and the allele frequency was estimated at every site by applying Eq. (1) in [42]:

$$\mathrm{P}(\mathrm{Z}_i\,|\,Y_i) = \prod_{j:Z_{i,j}=1}\left((1-\varepsilon)\frac{Y_i}{n}+\varepsilon\left(1-\frac{Y_i}{n}\right)\right) \times \prod_{j:Z_{i,j}=0}\left((1-\varepsilon)\left(1-\frac{Y_i}{n}\right)+\varepsilon\frac{Y_i}{n}\right) \tag{3}$$

where Y_i is the number of derived alleles at the *i*th genomic position in *n* sampled chromosomes, Z_i is the observed reads at the *i*th position with coverage r_i, $Z_\mathrm{i,j}$ $(1 \le j \le r_\mathrm{i})$ is an indicator variable equal to 1 if the *j*th read has the derived mutation, and 0 otherwise, and ε is the error rate of sequencing (Additional file 2: Text S1). Because we did not know the ancestral state of alleles, we used $P_\mathrm{f}(Z_\mathrm{i}|Y_\mathrm{i}) = 1/2P(Z_\mathrm{i}|Y_\mathrm{i}) + 1/2P(1\text{-}Z_\mathrm{i}|Y_\mathrm{i})$ to fold the SFS. We estimated the SFS as the probability is maximized. We also used the Eq. (3) to estimate an allele frequency at every site (Additional file 2: Text S1). A genome scan was performed to identify highly differentiated genomic regions. We performed a sliding window analysis in 10-kb windows with 5-kb increments after discarding windows, in which < 50% of sites were covered by ≥20× reads. We calculated the F_ST value in each window and screened windows with the top 0.1% of F_ST values. A neutrality test was performed using the ms software [76]. A coalescent simulation of 80 chromosomes was performed given the inferred population expansion (Additional file 2: Text S1). The length of the simulated region was set to 10 kb, which was the same as the window size used for the genome scan. The 80 chromosomes were randomly divided into subsamples of 40 to simulate a panmictic population. Then, Pool-seq data were simulated, the allele frequency was estimated for each site, and F_ST was calculated. These processes was repeated 100,000 times. To maximize the variance of the null distribution of F_ST, we assumed no recombination and set coverage to 20× for both species.

The effect of divergent selection under the pressure of migration

We assumed an isolation with migration model with two populations (populations 1 and 2) with population size, *N*, into which we incorporated mutation, recombination, migration, divergent selection, and random genetic drift. We consider a two-locus biallelic model. The locus I has *A* and *a* alleles and is the target of divergent selection, where the fitness values of A and a in population 1 are 1 and $1–s$, and those in population 2 are $1–s$ and 1, and s is the selection coefficient. We assume that divergent selection acts in an additive manner. The locus II has B and b alleles with no phenotypic effect. Symmetric mutation occurs only in locus II at rate μ per generation to measure the hitchhiking effect of divergent selection. Recombination between the two loci is incorporated at rate r per generation. Symmetric migration occurs at the rate m per gamete per generation between the populations. Let the frequencies of A-B, A-b, a-B, and a-b in population 1 be x_1, x_2, x_3, and x_4, respectively. As such, let those in population 2 be y_1, y_2, y_3, and y_4, respectively. The expectations of these frequencies in the next generation can be given by the following recursion equations:

$$\begin{aligned} x_1{}' = {} & (1-\mu)\, x_1 + \mu\, x_2 - r\, D_x - m\,(x_1 - y_1) \\ & + s\, x_1\,(x_3 + x_4) \end{aligned} \tag{3a}$$

$$\begin{aligned} x_2{}' = {} & (1-\mu)\, x_2 + \mu\, x_1 + r\, D_x - m\,(x_2 - y_2) \\ & + s\, x_2\,(x_3 + x_4) \end{aligned} \tag{3b}$$

$$\begin{aligned} x_3{}' = {} & (1-\mu)\, x_3 + \mu\, x_4 + r\, D_x - m\,(x_3 - y_3) \\ & - s\, x_3\,(x_1 + x_2) \end{aligned} \tag{3c}$$

$$\begin{aligned} x_4{}' = {} & (1-\mu)\, x_4 + \mu\, x_3 - r\, D_x - m\,(x_4 - y_4) \\ & - s\, x_4\,(x_1 + x_2) \end{aligned} \tag{3d}$$

$$\begin{aligned} y_1{}' = {} & (1-\mu)\, y_1 + \mu\, y_2 - r\, D_y - m\,(y_1 - x_1) \\ & - s\, y_1\,(y_3 + y_4) \end{aligned} \tag{3e}$$

$$\begin{aligned} y_2{}' = {} & (1-\mu)\, y_2 + \mu\, y_1 + r\, D_y - m\,(y_2 - x_2) \\ & - s\, y_2\,(y_3 + y_4) \end{aligned} \tag{3f}$$

$$\begin{aligned} y_3{}' = {} & (1-\mu)\, y_3 + \mu\, y_4 + r\, D_y - m\,(y_3 - x_3) \\ & + s\, y_3\,(y_1 + y_2) \end{aligned} \tag{3g}$$

$$\begin{aligned} y_4{}' = {} & (1-\mu)\, y_4 + \mu\, y_3 - r\, D_y - m\,(y_4 - x_4) \\ & + s\, y_4\,(y_1 + y_2) \end{aligned} \tag{3i}$$

where $D_x = x_1\, x_4 - x_2\, x_3$ and $D_y = y_1\, y_4 - y_2\, y_3$.

We simulated the pattern of polymorphisms around the target site of divergent selection. We fixed $N = 1000$, and $4N\mu = 0.01$. We used a wide range of $4Nr$ to be 0.1~200. We assumed that the locus I is monomorphic for A in both populations, and performed a pre-run until the DNA polymorphism in locus II reached a mutation-drift equilibrium using the Eq. (3a, b, c, d, e, f, g, h, i). At time $T = 0$, we introduced a in population 2 with the initial frequency, $1/2\,N$, and ran the simulation for $8\,N$ generations. We ran simulations for 100,000 cycles each with the variety of pairs of m and s, and calculated expected π_W, π_B, and F_{ST} in locus II at several time points.

RNA-seq and assembly

To screen the candidates of differentially expressed genes between the two species, a pooled RNA-seq analysis was performed. Total RNAs were extracted from the anterior part of the lateral skins of five males each from *H. pyrrhocephalus* and *H.* sp. 'macula.' Equal amounts of total RNAs (1 μg) were pooled, libraries were constructed using the TruSeq RNA Library Preparation Kit (Illumina), and the sequences were determined (paired-end 100 bp) using the Illumina HiSeq2000 platform. De novo assembly of paired-end short reads (7.7 Gbp) of *H.* sp. 'macula' was performed using the CLC genomic workbench (https://www.qiagenbioinformatics.com/) with automatic word size. The short reads from both species (*H. pyrrhocephalus*, 6.3 Gbp; *H.* sp. 'macula,' 7.7 Gbp) were mapped to the assembled sequences (50,240 contigs) and the expression levels of sequences were compared between species using the CLC genomic workbench. The sequences with different expression between species (*t*-test with Bonferroni correction, $P < 0.05$) were differentially expressed candidate genes. In total, 50,240 contigs were tested and 172 (0.3%) showed differential expression. For gene identification, the differentially expressed contig sequences were subjected to a BLASTN search against the NCBI non-redundant nucleotide sequences database (https://www.ncbi.nlm.nih.gov/). The differentially expressed contigs that were found in DRs were selected as candidate genes for differential expression between the two species.

Real-time qPCR

The expression of the candidate genes for differential expression screened by pooled RNA-seq were further analyzed by real-time qPCR (qPCR) between laboratory-reared individuals of *H.* sp. 'macula' and *H. pyrrhocephalus.* Fishes were 9–12 months old and were kept at 25 °C under commercial fluorescent lights with a 12 h light-dark cycle. To sample eye tissues, six individuals each from both species were euthanized under anesthesia using ethyl 4-aminobenzoate at 10 h after the light was turned on, and right eyes were enucleated. The eyes were immediately placed on ice in RNAlater (Ambion, Austin, TX, USA) and the cornea and lens were removed. The remaining eye samples were stored in fresh RNAlater at 80 °C until further use. To sample skin tissues, 10 individuals (five males and five females) each from both species were euthanized as described above at 5 h after the light was turned on. The euthanized fishes were immediately placed on ice in RNAlater and subsequent dissection was performed in this solution. A square area of the anterior part of the lateral skin was dissected. After muscle attached to the dissected skin was removed, the skin was cut into 2–5 mm^2 pieces. The skin pieces were stored in fresh RNAlater at 80 °C until further use.

Total RNA was extracted from the eye and skin samples using TRIzol RNA Isolation Reagent (Thermo Fisher Scientific, Waltham, MA, USA) according to the manufacturer's instructions and quantified using a NanoDrop 2000c spectrophotometer (Thermo Fisher Scientific). First-strand cDNA was reverse-transcribed from 500 ng of the eye total RNA or 1 μg of the skin total RNA using a PrimeScript RT Reagent Kit with gDNA Eraser (TaKaRa). The eye cDNA samples were diluted 25-fold in PCR-grade water for the amplification of melanopsin A. The skin cDNA samples were diluted 1.5- or 20-fold in PCR-grade water for the amplification of *aloxe3* and UDP-*N*-acetylglucosamine transporter or for *P450*, respectively. Target genes and an internal control gene (*GAPDH*) were amplified from the cDNA samples in a 25-μl total volume of PCR solution containing 12.5 ml of SYBR *Premix Ex Taq* II (TaKaRa), 3 ml of the diluted cDNA samples, and 10 pmol each of the following forward and reverse primers: melanopsin A: 5′ – TGGAGCTTTCATCGATGGCTACAAC– 3′ and 5′ – GATGCCTACAGCAAGGATGACAACAC– 3′; *GAPDH*: 5′ – GCCCACGCAAACATCATTC– 3′ and 5′ – GTCAGATCCACCACTGACACATC– 3′; *aloxe3*: 5′ – GAAGCTGCAAGGTGACAGGACTATTG– 3′ and 5′ – TGAGATGGTCAAGTTCGTCACCATG– 3′; *P450*: 5′ – GAGAAATCTGAACGCAGACTGCAAAC– 3′ and 5′ – GGACAGCAGTGACTTCTGATGCTCTATC– 3′; UDP-*N*-acetylglucosamine transporter: 5′ – AGCGAGG ACAGGACCATCAAGAG– 3′ and 5′ – GAGACACGT ATTTTAGCCTGGAGGAAAG– 3′. PCRs were performed using the Thermal Cycler Dice Real Time System II (TaKaRa) with the following conditions: 95 °C for 30 s, followed by 40 cycles of 95 °C for 5 s and 60 °C for 30 s. Correction of the PCR efficiency for each primer set was performed using a standard curve drawn from the dilution series of the cDNA samples. *GAPDH* was used as an internal control. Each sample was measured at least two times for technical replicates.

DR sequence determination and phylogenetic tree construction

To confirm the fixed differences in differentiated regions (DRs), sets of primers were designed for four DRs (DR11, DR12, DR17, and DR19) to amplify regions including fixed differences. The primer sequences are listed below. The primers for DR5 were reported previously [15, 16]. Five DRs were amplified by PCR with the following conditions: a denaturation step for 3 min at 94 °C followed by 30 cycles of denaturation for 1 min at 94 °C, annealing for 30 s at 55 °C, and extension for 30 s at 72 °C. PCR products were purified and the sequences were determined using the Applied Biosystems Automated 3130xl Sequencer (Applied Biosystems, Waltham, MA, USA).

To construct phylogenetic trees, the sequences of DRs (~ 1 kb) and orthologous sequences from genome sequence data were used. Sets of primers were designed for 16 DRs (the remaining DRs failed to amplify), and these were amplified by PCR with the following reaction conditions: a denaturation step for 3 min at 94 °C followed by 30 cycles of denaturation for 1 min at 94 °C, annealing for 30 s at 55 °C, and extension for 1.5 min at 72 °C. The primer sequences are listed below. PCR products were cloned into the T-Vector pMD20 vector (Takara, Shiga, Japan) and the sequences were determined using the Applied Biosystems Automated 3130xl Sequencer. The genomic DNAs used as templates for amplification were the Lake Victoria species *H. pyrrhocephalus*, *H.* sp. 'macula,' and *H. piceatus*; riverine species *H.* sp. 'katonga,' *H.* sp. 'kitilda-rukwa,' and *H.* sp. 'muzu' (see Additional file 4: Figure S4F for localities); Lake Malawi species *Labidochromis caeruleus*, *Melanochromis auratus*, *Labeotropheus trewavasae*, *Pseudotropheus lombardoi*, and *Dimidiochromis strigatus*; and Lake Tanganyika species *Tropheus moorii*, *T. duboisi*, *T. brichardi*, *Simochromis pleurospilus*, *Petrochromis macrognathus*, *Cyprichromis coloratus*, *Ectodus descampsi*, *Perissodus eccentricus*, and *Neolamprologus tretocephalus*. The consensus sequences of the genomes of *H. pyrrhocephalus* and *H.* sp. 'macula' were constructed from the mapping results of the paired-end short reads to the reference genome sequence of *P. nyererei* [14, 76] using the CLC genomic workbench. Orthologous sequences of DR sequences were obtained by BLASTN searches [77] and from genome sequence data for *Pundamilia nyererei*, *Metriaclima zebra*, *Astatotilapia burtoni*, *Neolamprologus brichardi*, and *Oreochromis niloticus* [14], and consensus sequences of *H. pyrrhocephalus* and *H.* sp. 'macula.' In the case of DR5, upstream (2 kbp) and downstream (2 kbp) sequences of *LWS* were determined following methods described in previous studies [15, 18] and using sequences from previous studies [15, 17, 18] deposited in databases. Each of the DR sequences was aligned and subjected to a phylogenetic analysis using the maximum likelihood method with 1000 bootstrap replications in MEGA ver. 6 [78].

Gene ontology analysis

The sequences of DRs (14–28 kbp) were used as queries for BLASTN searches [77] against the NCBI nucleotide database (http://blast.ncbi.nlm.nih.gov). The sequences of genes in DRs were subjected to a gene ontology analysis using DAVID [79] and Blast2GO [80]. The details of the Gene Ontology Analysis and primer sequences can be found in Supplemental Material online.

Additional files

Additional file 1: Figure S1. (A) Site frequency spectrum of Lake Victoria cichlids. The white and black bars represent *H. pyrrhocephalus* and *H.* sp. 'macula,' respectively. (B) Demographic model of Lake Victoria cichlids. See the section Demographic Model and Parameter Estimation for details. (C) Nucleotide and indel frequencies within five DRs. We amplified and determined the sequences, including fixed differences, of four DRs from 20 individuals each of *H.* sp. 'macula' and *H. pyrrhocephalus*. Positions indicate the positions from the first nucleotides of the determined sequences. The frequencies of nucleotides in the coding region of *LWS* were verified in a previous study (*14*). (PDF 155 kb)

Additional file 2 Text S1. Supporting Text. (DOCX 572 kb)

Additional file 3 Figure S2. Population genomic analyses using only male individuals. (A) Site frequency spectrum in *H. pyrrhocephalus* (white bars; 126,843 SNPs) and *H.* sp. 'macula' (black bars; 146,456 SNPs). Nucleotide diversity and Tajima's *D* values were almost same as those in all samples. (B) Average *FST* values (±1 SD) against coverage for Pool-seq data as in Fig. 1b (5,662,990 SNPs). The blue and orange dots represent observed and simulated values under panmixia. (C) The spatial patterns of average nucleotide diversity within each species (πW; pink), average pairwise nucleotide divergence between species (πB; blue), and *FST* (green) in and around *LWS* gene as in Fig. 2b. The green arrows represent fixed nucleotide differences between species. (PDF 279 kb)

Additional file 4: Figure S3. The spatial patterns of average nucleotide diversity within species (πW; pink), average pairwise nucleotide divergence between species (πB; blue), and *FST* (green) in and around DRs. The green arrows represent fixed nucleotide differences between species. (PDF 1356 kb)

Additional file 5: Figure S4. The origins of mutations in DRs. Three phylogenetic trees represent the accumulation of mutations in the common ancestral species of (A) Lake Victoria species, (B) Lake Victoria and riverine *Haplochromis* species, and (C) tribe Tropheini in Lake Tanganyika, Lakes Malawi, Victoria, and riverine *Haplochromis* species. The tree topologies constructed from sequences of each DR were consistent with (D) "Riverine origin" or (E) "Modern haplochromine origin." Scale bars indicate the number of substitutions per site. (F) The LWS sequences were determined from three riverine species: *H.* sp. 'katonga' from Katonga, *H.* sp. 'kitilda-rukwa' from Kitilda-Rukwa, and *H.* sp. 'muzu' from Muzu. (PDF 554 kb)

Additional file 6: Table S1. Genes in DRs. (PDF 149 kb)

Abbreviation

SFS: site frequency spectrum

Acknowledgments

We thank S. Wright and J. Gojobori for providing helpful comments on the early version of our manuscript, and the Tanzania Fisheries Research Institute (TAFIRI) for help with sampling.

Funding

This work was supported by KAKENHI Grant Number JP23570269 and JP26440209 to Y.T., and an internal SOKENDAI grant to Y.T.

Availability of data and materials

The nucleotide sequences were deposited in GenBank under accession numbers LC129373–LC129499 and in the DDBJ Sequenced Read Archive under accession numbers DRX051884–DRX051889.

Authors' contributions

ST performed the next-generation sequence data analysis, population genomic analysis, determination of DRs and long DRs, and manuscript writing. RM developed the research concept, and performed DNA and RNA extraction, quantitative PCR analysis, the identification of genes in DRs, and manuscript writing. JO performed gene ontology analyses for genes in DRs and long DRs. STK determined DR sequences and performed phylogenetic analyses. AS managed riverine *Haplochromis* species samples and analyzed DR sequences. HT performed sampling and identification of riverine *Haplochromis* species. MN provided helpful discussion. MA performed sampling and identification of Lake Victoria species. SM performed sampling and identification of Lake Victoria species. HDJM performed sampling of Lake Victoria species. SM was involved in management and sampling of Lake Victoria species. NO arranged the sampling of Lake Victoria species. YT developed the research concept, planned the research, and performed DNA and RNA extraction, mapping of short reads, identification of genes in DRs and long DRs, determination of DR sequences for phylogenetic analyses, and manuscript writing. All authors read and approved the final manuscript.

Ethics approval and consent to participate

The animal protocols and procedures were approved by the Institutional Animal Care and Use Committee of Tokyo Institute of Technology.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Author details

[1]Department of Evolutionary Studies of Biosystems, SOKENDAI (The Graduate University for Advanced Studies), Shonan Village, Hayama, Kanagawa 240-0193, Japan. [2]Graduate School of Bioscience and Biotechnology, Tokyo Institute of Technology, 4259 Nagatsuta-cho, Midori-ku, Yokohama 226-8501, Japan. [3]Department of Biological sciences, Tokyo Metropolitan University, 1-1 Minamiosawa, Hachioji, Tokyo 197-0397, Japan. [4]JST (Japan Science and Technology Agency), NBDC (National Bioscience Database Center), 5-3, Yonbancho, Chiyoda-ku, Tokyo 102-0081, Japan. [5]Department of Anatomy and Cytohistology, School of Dental Medicine, Tsurumi University, 2-1-3 Tsurumi, Tsurumi-ku, Yokohama 230-8501, Japan. [6]Max-Planck-Institut für Biologie, Abteilung Immungenetik, Corrensstrasse 42, D-72076 Tübingen, Germany. [7]School of Life Science and Technology, Department of Life Science and Technology, Tokyo Institute of Technology (Tokyo Tech), 2-12-1, Ookayama, Meguro ward, Tokyo, Japan. [8]Tanzania Fisheries Research Institute (TAFIRI), Mwanza, Tanzania. [9]Department of Life Sciences, National Cheng Kung University, 701 Tainan, Taiwan. [10]Foundation for Advancement of International Science (FAIS), Tsukuba, Japan.

Received: 28 November 2017 Accepted: 12 February 2019
Published online: 04 March 2019

References

1. Coyne JA, Orr HA: Speciation, vol. 37: Sinauer Associates Sunderland, MA; 2004.
2. Kocher TD. Adaptive evolution and explosive speciation: the cichlid fish model. Nat Rev Genet. 2004;5(4):288–98.
3. Seehausen O. Lake Victoria rock cichlids: taxonomy, ecology and distribution: Verduyn cichlids; 1996.
4. Turner GF, Seehausen O, Knight ME, Allender CJ, Robinson RL. How many species of cichlid fishes are there in African lakes? Mol Ecol. 2001;10(3):793–806.
5. Johnson TC, Kelts K, Odada E. The holocene history of Lake Victoria. AMBIO: A Journal of the Human Environment. 2000;29(1):2–11.
6. Johnson TC, Scholz CA, Talbot MR, Kelts K, Ricketts RD, Ngobi G, Beuning K, Ssemmanda I, McGill JW. Late Pleistocene desiccation of Lake Victoria and rapid evolution of cichlid fishes. Science. 1996;273(5278):1091–3.
7. Meier JI, Sousa VC, Marques DA, Selz OM, Wagner CE, Excoffier L, Seehausen O. Demographic modelling with whole-genome data reveals parallel origin of similar Pundamilia cichlid species after hybridization. Mol Ecol. 2017;26(1):123–41.
8. Meier JI, Marques DA, Wagner CE, Excoffier L, Seehausen O. Genomics of parallel ecological speciation in Lake Victoria cichlids. Mol Biol Evol. 2018;35(6):1489–506.
9. Nagl S, Tichy H, Mayer WE, Takahata N, Klein J. Persistence of neutral polymorphisms in Lake Victoria cichlid fish. Proc Natl Acad Sci. 1998;95(24):14238–43.

10. Terai Y, Takezaki N, Mayer WE, Tichy H, Takahata N, Klein J, Okada N. Phylogenetic relationships among east African haplochromine fish as revealed by short interspersed elements (SINEs). J Mol Evol. 2004;58(1):64–78.
11. Samonte IE, Satta Y, Sato A, Tichy H, Takahata N, Klein J. Gene flow between species of Lake Victoria haplochromine fishes. Mol Biol Evol. 2007;24(9):2069–80.
12. Maeda K, Takeda M, Kamiya K, Aibara M, Mzighani SI, Nishida M, Mizoiri S, Sato T, Terai Y, Okada N. Population structure of two closely related pelagic cichlids in Lake Victoria, Haplochromis pyrrhocephalus and H. Laparogramma. Gene. 2009;441(1):67–73.
13. Wagner CE, Keller I, Wittwer S, Selz OM, Mwaiko S, Greuter L, Sivasundar A, Seehausen O. Genome-wide RAD sequence data provide unprecedented resolution of species boundaries and relationships in the Lake Victoria cichlid adaptive radiation. Mol Ecol. 2013;22(3):787–98.
14. Brawand D, Wagner CE, Li YI, Malinsky M, Keller I, Fan S, Simakov O, Ng AY, Lim ZW, Bezault E, et al. The genomic substrate for adaptive radiation in African cichlid fish. Nature. 2014;513(7518):375–81.
15. Terai Y, Seehausen O, Sasaki T, Takahashi K, Mizoiri S, Sugawara T, Sato T, Watanabe M, Konijnendijk N, Mrosso HD, et al. Divergent selection on opsins drives incipient speciation in Lake Victoria cichlids. PLoS Biol. 2006;4(12):e433.
16. Terai Y, Mayer WE, Klein J, Tichy H, Okada N. The effect of selection on a long wavelength-sensitive (LWS) opsin gene of Lake Victoria cichlid fishes. Proc Natl Acad Sci U S A. 2002;99(24):15501–6.
17. Seehausen O, Terai Y, Magalhaes IS, Carleton KL, Mrosso HD, Miyagi R, van der Sluijs I, Schneider MV, Maan ME, Tachida H, et al. Speciation through sensory drive in cichlid fish. Nature. 2008;455(7213):620–6.
18. Miyagi R, Terai Y, Aibara M, Sugawara T, Imai H, Tachida H, Mzighani SI, Okitsu T, Wada A, Okada N. Correlation between nuptial colors and visual sensitivities tuned by opsins leads to species richness in sympatric Lake Victoria cichlid fishes. Mol Biol Evol. 2012;29(11):3281–96.
19. Terai Y, Miyagi R, Aibara M, Mizoiri S, Imai H, Okitsu T, Wada A, Takahashi-Kariyazono S, Sato A, Tichy H, et al. Visual adaptation in Lake Victoria cichlid fishes: depth-related variation of color and scotopic opsins in species from sand/mud bottoms. BMC Evol Biol. 2017;17(1):200.
20. Meier JI, Marques DA, Mwaiko S, Wagner CE, Excoffier L, Seehausen O. Ancient hybridization fuels rapid cichlid fish adaptive radiations. Nat Commun. 2017;8:14363.
21. Wu CI. The genic view of the process of speciation. J Evol Biol. 2001;14(6): 851–65.
22. Innan H, Watanabe H. The effect of gene flow on the coalescent time in the human-chimpanzee ancestral population. Mol Biol Evol. 2006;23(5):1040–7.
23. Wolf JB, Ellegren H. Making sense of genomic islands of differentiation in light of speciation. Nat Rev Genet. 2016.
24. Feder JL, Egan SP, Nosil P. The genomics of speciation-with-gene-flow. Trends Genet. 2012;28(7):342–50.
25. Papadopulos AS, Baker WJ, Crayn D, Butlin RK, Kynast RG, Hutton I, Savolainen V. Speciation with gene flow on Lord Howe Island. Proc Natl Acad Sci. 2011;108(32):13188–93.
26. Martin SH, Dasmahapatra KK, Nadeau NJ, Salazar C, Walters JR, Simpson F, Blaxter M, Manica A, Mallet J, Jiggins CD. Genome-wide evidence for speciation with gene flow in Heliconius butterflies. Genome Res. 2013;23(11):1817–28.
27. Nadeau NJ, Martin SH, Kozak KM, Salazar C, Dasmahapatra KK, Davey JW, Baxter SW, Blaxter ML, Mallet J, Jiggins CD. Genome-wide patterns of divergence and gene flow across a butterfly radiation. Mol Ecol. 2013;22(3):814–26.
28. Malinsky M, Challis RJ, Tyers AM, Schiffels S, Terai Y, Ngatunga BP, Miska EA, Durbin R, Genner MJ, Turner GF. Genomic islands of speciation separate cichlid ecomorphs in an east African crater lake. Science. 2015;350(6267):1493–8.
29. Malinsky M, Svardal H, Tyers AM, Miska EA, Genner MJ, Turner GF, Durbin R. Whole genome sequences of Malawi cichlids reveal multiple radiations interconnected by gene flow. Nature Ecology & Evolution. 2018;2:1940–55.
30. Poelstra JW, Vijay N, Bossu CM, Lantz H, Ryll B, Muller I, Baglione V, Unneberg P, Wikelski M, Grabherr MG, et al. The genomic landscape underlying phenotypic integrity in the face of gene flow in crows. Science. 2014;344(6190):1410–4.
31. Lamichhaney S, Berglund J, Almen MS, Maqbool K, Grabherr M, Martinez-Barrio A, Promerova M, Rubin CJ, Wang C, Zamani N, et al. Evolution of Darwin's finches and their beaks revealed by genome sequencing. Nature. 2015;518(7539):371–5.
32. Turner TL, Hahn MW, Nuzhdin SV. Genomic islands of speciation in Anopheles gambiae. PLoS Biol. 2005;3(9):e285.
33. Ellegren H, Smeds L, Burri R, Olason PI, Backstrom N, Kawakami T, Kunstner A, Makinen H, Nadachowska-Brzyska K, Qvarnstrom A, et al. The genomic landscape of species divergence in Ficedula flycatchers. Nature. 2012;491(7426):756–60.
34. Soria-Carrasco V, Gompert Z, Comeault AA, Farkas TE, Parchman TL, Johnston JS, Buerkle CA, Feder JL, Bast J, Schwander T. Stick insect genomes reveal natural selection's role in parallel speciation. Science. 2014;344(6185):738–42.
35. Goldschmidt T, Witte F, De Visser J. Ecological segregation in zooplanktivorous haplochromine species (Pisces: Cichlidae) from Lake Victoria. Oikos. 1990:343–55.
36. Witte F. Ecological differentiation in Lake Victoria haplochromines: comparison of cichlid species flocks in African Lakes; 1984.
37. Witte F, Goldschmidt T, Wanink J, van Oijen M, Goudswaard K, Witte-Maas E, Bouton N. The destruction of an endemic species flock: quantitative data on the decline of the haplochromine cichlids of Lake Victoria. Environ Biol Fish. 1992;34(1):1–28.
38. Greenwood PH. A revision of the Lake Victoria Haplochromis species (Pisces, Cichlidae): part 4: British museum natural history; 1960.
39. Fryer G, Iles TD. The cichlid fishes of the great lakes of Africa: their biology and evolution: Oliver and Boyd; 1972.
40. Seehausen O, Van Alphen JJ, Witte F. Cichlid fish diversity threatened by eutrophication that curbs sexual selection. Science. 1997;277(5333):1808–11.
41. Seehausen O, van Alphen JJ. The effect of male coloration on female mate choice in closely related Lake Victoria cichlids (Haplochromis nyererei complex). Behav Ecol Sociobiol. 1998;42(1):1–8.
42. Boitard S, Schlotterer C, Nolte V, Pandey RV, Futschik A. Detecting selective sweeps from pooled next-generation sequencing samples. Mol Biol Evol. 2012;29(9):2177–86.
43. Tajima F. Evolutionary relationship of DNA sequences in finite populations. Genetics. 1983;105(2):437–60.
44. Tajima F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics. 1989;123(3):585–95.
45. Slatkin M, Hudson RR. Pairwise comparisons of mitochondrial DNA sequences in stable and exponentially growing populations. Genetics. 1991;129(2):555–62.
46. Gutenkunst RN, Hernandez RD, Williamson SH, Bustamante CD. Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. PLoS Genet. 2009;5(10):e1000695.
47. Lynch M, Bost D, Wilson S, Maruki T, Harrison S. Population-genetic inference from pooled-sequencing data. Genome biology and evolution. 2014;6(5):1210–8.
48. Takeda M, Kusumi J, Mizoiri S, Aibara M, Mzighani SI, Sato T, Terai Y, Okada N, Tachida H. Genetic structure of pelagic and littoral cichlid fishes from Lake Victoria. PLoS One. 2013;8(9):e74088.
49. Hudson RR, Slatkin M, Maddison W. Estimation of levels of gene flow from DNA sequence data. Genetics. 1992;132(2):583–9.
50. Hein J, Schierup M, Wiuf C: Gene genealogies, variation and evolution: a primer in coalescent theory: Oxford University press, USA; 2004.
51. Smith JM, Haigh J. The hitch-hiking effect of a favourable gene. Genet Res. 1974;23(1):23–35.
52. Kim Y, Stephan W. Detecting a local signature of genetic hitchhiking along a recombining chromosome. Genetics. 2002;160(2):765–77.
53. Teshima KM, Innan H. Neofunctionalization of duplicated genes under the pressure of gene conversion. Genetics. 2008.
54. Cruickshank TE, Hahn MW. Reanalysis suggests that genomic islands of speciation are due to reduced diversity, not reduced gene flow. Mol Ecol. 2014;23(13):3133–57.
55. Wolf JB, Ellegren H. Making sense of genomic islands of differentiation in light of speciation. Nat Rev Genet. 2017;18(2):87.
56. Nordborg M, Charlesworth B, Charlesworth D. The effect of recombination on background selection. Genet Res. 1996;67(2):159–74.
57. Salzburger W, Mack T, Verheyen E, Meyer A. Out of Tanganyika: genesis, explosive speciation, key-innovations and phylogeography of the haplochromine cichlid fishes. BMC Evol Biol. 2005;5(1):17.
58. Zhang Q, Eisenstat DD. Roles of homeobox genes in retinal ganglion cell differentiation and axonal guidance. Adv Exp Med Biol. 2012;723:685–91.
59. Hankins MW, Peirson SN, Foster RG. Melanopsin: an exciting photopigment. Trends Neurosci. 2008;31(1):27–36.

60. Van der Meer H, Bowmaker J. Interspecific variation of photoreceptors in four co-existing haplochromine cichlid fishes. Brain Behav Evol. 1995;45(4):232–40.
61. Fernald RD, Maruska KP. Social information changes the brain. Proc Natl Acad Sci U S A. 2012;109(Suppl 2):17194–9.
62. Piferrer F, Blázquez M. Aromatase distribution and regulation in fish. Fish Physiol Biochem. 2005;31(2–3):215–26.
63. Nielsen EE, Hemmer-Hansen J, Poulsen NA, Loeschcke V, Moen T, Johansen T, Mittelholzer C, Taranger GL, Ogden R, Carvalho GR. Genomic signatures of local directional selection in a high gene flow marine organism; the Atlantic cod (Gadus morhua). BMC Evol Biol. 2009;9:276.
64. Hecky R, Bugenyi F, Ochumba P, Talling J, Mugidde R, Gophen M, Kaufman L. Deoxygenation of the deep water of Lake Victoria, East Africa. Limnol Oceanogr. 1994;39(6):1476–81.
65. Rutjes HA, Nieveen MC, Weber RE, Witte F, Van den Thillart GE. Multiple strategies of Lake Victoria cichlids to cope with lifelong hypoxia include hemoglobin switching. Am J Physiol Regul Integr Comp Physiol. 2007;293(3):R1376–83.
66. Chapman LJ, Kaufman LS, Chapman CA, McKenzie FE. Hypoxia tolerance in twelve species of east African cichlids: potential for low oxygen refugia in Lake Victoria. Conserv Biol. 1995;9(5):1274–88.
67. Etzold S, Juge N. Structural insights into bacterial recognition of intestinal mucins. Curr Opin Struct Biol. 2014;28:23–31.
68. Baumgart M, Groth M, Priebe S, Savino A, Testa G, Dix A, Ripa R, Spallotta F, Gaetano C, Ori M. RNA-seq of the aging brain in the short-lived fish N. Furzeri–conserved pathways and novel genes associated with neurogenesis. Aging Cell. 2014;13(6):965–74.
69. Loh YH, Bezault E, Muenzel FM, Roberts RB, Swofford R, Barluenga M, Kidd CE, Howe AE, Di Palma F, Lindblad-Toh K, et al. Origins of shared genetic variation in African cichlids. Mol Biol Evol. 2013;30(4):906–17.
70. Witte F, Witte-Maas E: Implications for taxonomy and functional morphology of intraspecific variation in haplochromine cichlids of Lake Victoria. *Witte F, From form to fishery, PhD Thesis, Leiden, the Netherlands: Leiden University* 1987:1–83.
71. Witte F, Van Oijen M. Taxonomy, ecology and fishery of Lake Victoria haplichromine trophic groups: Nationaal Natuurhistorisch museum; 1990.
72. Mzighani SI, Nikaido M, Takeda M, Seehausen O, Budeba YL, Ngatunga BP, Katunzi EF, Aibara M, Mizoiri S, Sato T. Genetic variation and demographic history of the Haplochromis laparogramma group of Lake Victoria—an analysis based on SINEs and mitochondrial DNA. Gene. 2010;450(1):39–47.
73. Langmead B, Salzberg SL. Fast gapped-read alignment with bowtie 2. Nat Methods. 2012;9(4):357–9.
74. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. The sequence alignment/map format and SAMtools. Bioinformatics. 2009;25(16):2078–9.
75. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST. The variant call format and VCFtools. Bioinformatics. 2011;27(15):2156–8.
76. Hudson RR. Generating samples under a Wright–fisher neutral model of genetic variation. Bioinformatics. 2002;18(2):337–8.
77. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990;215(3):403–10.
78. Tamura K, Stecher G, Peterson D, Filipski A, Kumar S: MEGA6: molecular evolutionary genetics analysis version 6.0. *Molecular biology and evolution* 2013:mst197.
79. Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc. 2009;4(1):44–57.
80. Conesa A, Götz S, García-Gómez JM, Terol J, Talón M, Robles M. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics. 2005;21(18):3674–6.