Science AAAS

Accepted Manuscript

Title: Genomic islands of speciation separate cichlid ecomorphs in an East African crater lake[*]

Authors: Milan Malinsky[1,2]†, Richard J. Challis[3]†‡, Alexandra M. Tyers[3], Stephan Schiffels[1], Yohey Terai[4], Benjamin P. Ngatunga[5], Eric A. Miska[1,2], Richard Durbin[1], Martin J. Genner[6]*, George F. Turner[3]*

Affiliations:

[1]Wellcome Trust Sanger Institute, Cambridge, CB10 1SA, UK.

[2]Gurdon Institute and Department of Genetics, University of Cambridge, Cambridge, CB2 1QN, UK.

[3]School of Biological Sciences, Bangor University, Bangor, Gwynedd, LL57 2UW, UK.

[4]Department of Evolutionary Studies of Biosystems, SOKENDAI, Kanagawa 240-0193, Japan.

[5]Tanzania Fisheries Research Institute, Box 9750, Dar es Salaam, Tanzania.

[6]School of Biological Sciences, Life Sciences Building, 24 Tyndall Avenue, University of Bristol, Bristol, BS8 1TQ, UK.

*Corresponding authors. E-mail: george.turner@bangor.ac.uk, M.Genner@bristol.ac.uk.

†These authors contributed equally to this work.

‡Present address: Institute of Evolutionary Biology, University of Edinburgh, Charlotte Auerbach Road, Edinburgh EH9 3FL, UK

Abstract: The genomic causes and effects of divergent ecological selection during speciation are still poorly understood. Here, we report the discovery and detailed characterization of early-stage adaptive divergence of two cichlid fish ecomorphs in a small (700m diameter) isolated crater lake in Tanzania. The ecomorphs differ in depth preference, male breeding color, body shape, diet and trophic morphology. With whole genome sequences of 146 fish, we identify 98 clearly demarcated genomic 'islands' of high differentiation and demonstrate association of genotypes across these islands to divergent mate preferences. The islands contain candidate adaptive genes enriched for functions in sensory perception (including rhodopsin and other twilight vision associated genes), hormone signaling and morphogenesis. Our study suggests mechanisms and genomic regions that may play a role in the closely related mega-radiation of Lake Malawi.

One Sentence Summary: We describe the discovery of a pair of incipient species of African cichlid fish in a small isolated crater lake, and characterize their ecological, morphological and genomic separation, showing association of divergent genomic islands to mate choice.

Main Text:

Introduction

Understanding the causes and consequences of speciation, including at the genetic level, requires investigation of taxa at different stages on the speciation continuum (*1*, *2*). East African cichlids have repeatedly undergone rapid adaptive radiation (*3*). The Lake Malawi radiation has generated over 500 species in less than five million years, involving divergence in habitat,

[*] This manuscript has been accepted for publication in Science. This version has not undergone final editing. Please refer to the complete version of record at http://www.sciencemag.org/. The manuscript may not be reproduced or used in any manner that does not fall within the fair use provisions of the Copyright Act without the prior, written permission of AAAS.

feeding apparatus, and nuptial color. Thus these phenomena present an opportunity to observe hundreds of varied, recent, and in some cases ongoing, speciation events (*4*). However, investigation of early stages of speciation in the large cichlid radiations of Malawi as well as Lakes Tanganyika and Victoria has been hampered by difficulties in identifying sister species relationships, in reconstructing past geographical situations, and in controlling for possible introgression from non-sister taxa (*5*).

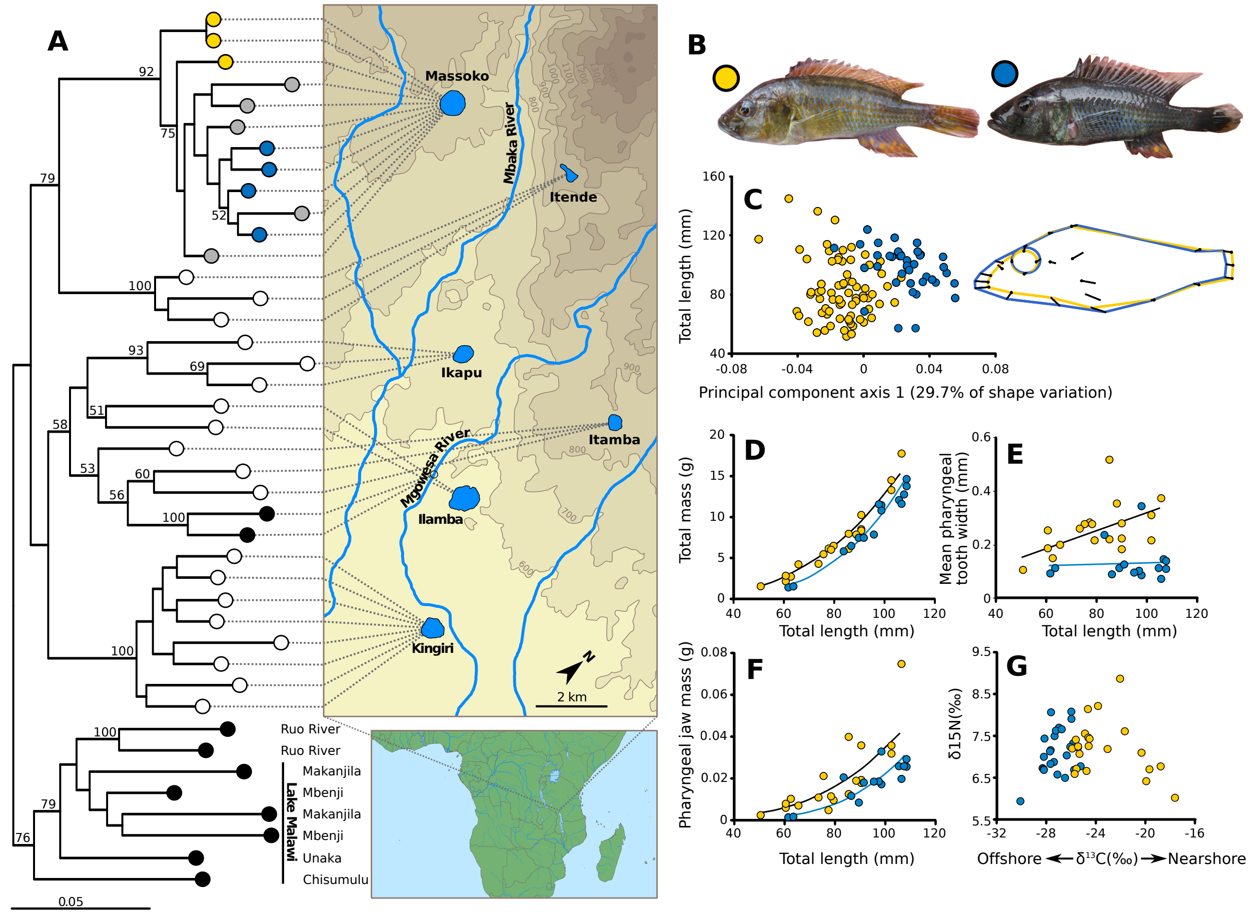


Fig. 1: Cichlid radiation in the crater lakes of southern Tanzania (A) A phylogeny of the crater lake *Astatotilapia* based on reference-aligned RAD data (7,906 SNPs across 5010 polymorphic RAD loci). It demonstrates reciprocal monophyly between the populations in each lake except for Itamba, and close relationship to *A. calliptera* from rivers and from Lake Malawi. Within Lake Massoko, yellow symbols indicate the littoral morph, blue symbols indicate the benthic, and grey symbols denote small, phenotypically ambiguous, and thus unassigned individuals. Additional *Astatotilapia* individuals from other crater lakes are denoted by open circles. *A. calliptera* from rivers and Lake Malawi are denoted by black circles. Bootstrap values are displayed for nodes with >50% support. **(B)** Breeding males of the yellow littoral and blue benthic morphs of Lake Massoko. The symbols next to the photographs correspond to symbols used in (C-G). **(C-F)** Morphological divergence between the two morphs of Lake Massoko. Relative to the littoral, the benthic morph has relatively longer head and jaw **(C),** lower body mass **(D)**, narrower 'papilliform' pharyngeal teeth **(E)**, and lighter lower pharyngeal jaws **(F)**. The benthic fish have stable isotope ratios that tend to be more depleted in C^{13} than the littoral, indicative of a more offshore-planktonic diet **(G)**.

During 2011, we conducted a survey (Table S1) of fish fauna in six crater lakes in the Rungwe District of Tanzania (Fig. 1A; Table S2). In all six lakes, we found endemic haplochromine cichlids of the genus *Astatotilapia,* closely related to *Astatotilapia calliptera* (Fig. 1A), a species widely distributed in the rivers, streams and shallow lake margins of the region. Thus, the Rungwe District *Astatotilapia* are close relatives of the of Lake Malawi endemic haplochromine cichlids (*5*).

In Lake Massoko (Fig. S1), the benthic zone in deep waters (~20-25m) is very dimly lit and populated by cichlids with phenotypes clearly different to those typical of shallow waters (~<5m) close to the shore (littoral). Deep-water males are dark blue-black, while most males collected from the shallow waters are yellow-green, similar to riverine *A. calliptera* (Fig. 1B; Movie S1; Table S3). We also collected small (<65mm standard length) males that were not readily field-assigned to either ecomorph (*6*). The benthic and littoral morphs are reminiscent of the species pair of *Pundamilia* cichlids from Lake Victoria (*7*), but within a potentially simpler historical and geographical context. Lake Massoko is steep-sided, has a strong thermocline at ~15m, and an anoxic boundary at ~25m (*8*). The estimated time of lake formation is ~50,000 years ago (*9*).

Ecomorph separation

To examine relationships between crater lake and riverine *A. calliptera* of southern Tanzania, we obtained restriction site associated DNA (RAD) data from 30 fish from the Rungwe District, and 11 outgroup *Astatotilapia* from the broader Lake Malawi catchment (Fig. S2, Table S4). A maximum likelihood phylogeny constructed on the basis of these data (*6*) demonstrates monophyly of all specimens from Lake Massoko (Fig. 1A). Thus, the RAD phylogeny provides evidence that Massoko morphs might have evolved in primary sympatry, as proposed for crater lake cichlid radiations of Cameroon (*10*) and Nicaragua (*11*).

Morphological analyses of these two color morphs revealed significant differences in head and body shape, body mass, the shape of pharyngeal teeth, and pharyngeal jaw mass (Fig. 1C-F; Table S5; ANCOVA tests, all $P<0.001$). We also found significant differences in stable isotope ratios (Fig. 1G; Table S5; ANCOVA test, $P<0.001$), indicative of dietary differences. Together these results demonstrate ecomorph separation and adaptation to different ecological environments.

Whole-genome evidence

To study the genome-wide pattern of Massoko ecomorph divergence and to further clarify its geographical context, we obtained whole-genome sequence data at ~15X coverage for 6 individuals each of the yellow littoral and blue benthic ecomorphs and 16 additional *A. calliptera* from the wider Lake Malawi catchment (Fig. S2), supplemented by lower coverage (~6X) data from 87 specimens from Lake Massoko (25 littoral, 32 benthic, and 30 small unassigned) and 30 individuals from Lake Itamba (Fig. 1A; Table S6). Sequence data were aligned to the *Metriaclima zebra* reference assembly (*12*), from which divergence was 0.2-0.3%, and variants were called at 4,755,448 sites (1.2-1.6 million sites per individual).

A maximum likelihood phylogeny built from whole genome sequence data confirmed reciprocal monophyly of *Astatotilapia* within Lakes Massoko and Itamba, and revealed the sister group of Massoko fish to be an *A. calliptera* population from the nearby Mbaka river (Fig. 2A). All specimens of the benthic ecomorph formed a monophyletic clade derived from the littoral ecomorph (Fig. 2A). Principal component analysis (PCA) showed strong population structure

(Tracy-Widom statistics: $P<1\times10^{-12}$), with benthic and littoral individuals separated by the first eigenvector and forming separate clusters (Fig. 2B). In contrast, within Lake Itamba, PCA did not reveal significant population structure (Tracy-Widom statistics: P=0.11). Individuals from Massoko that were not field-assigned to either of the ecomorphs did not form a monophyletic clade in the phylogeny (Fig. 2A) or a distinct cluster in PCA (Fig. 2B).

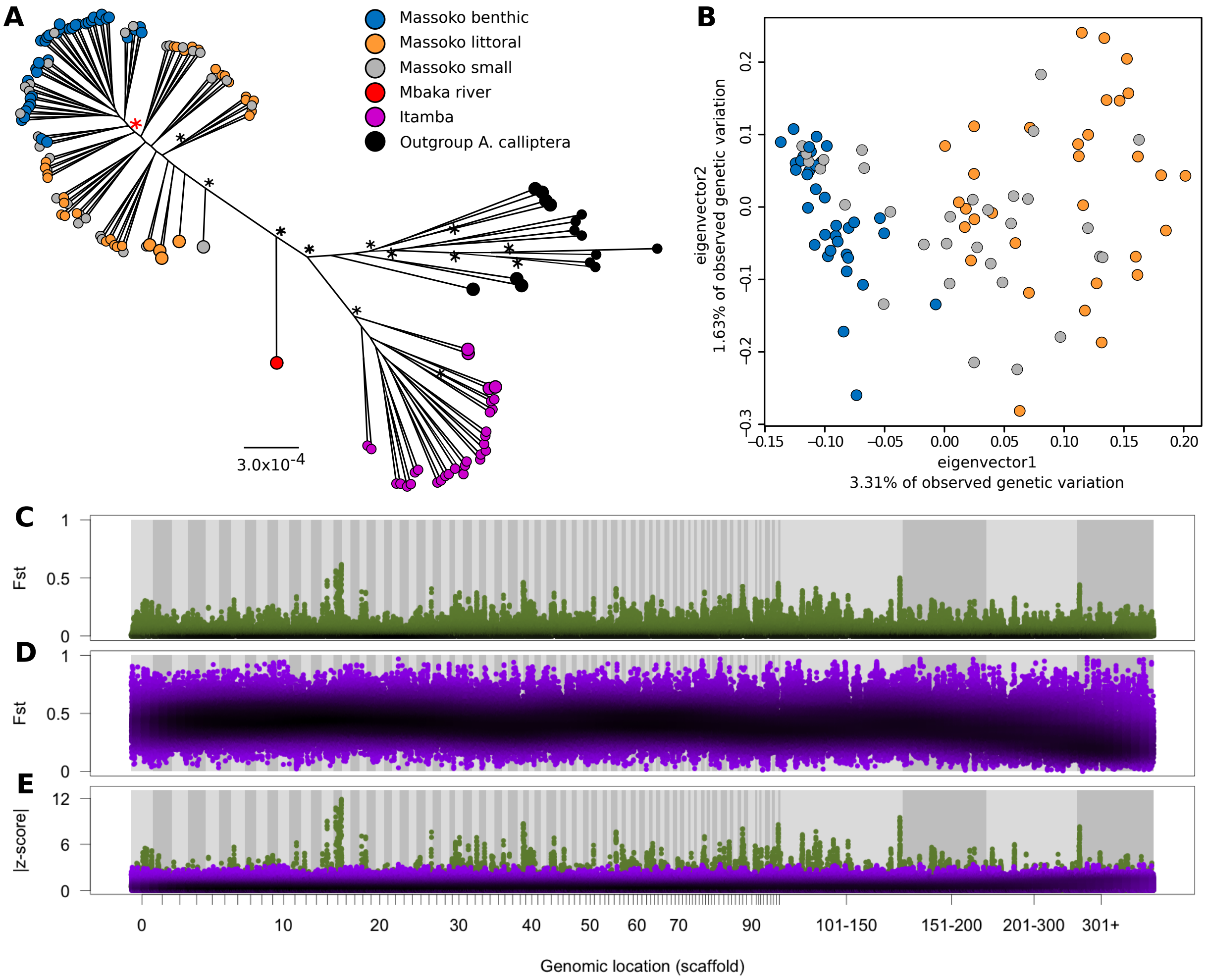


Fig. 2: Whole genome sequence data (A) A maximum likelihood whole-genome phylogenetic tree. Black stars indicate nodes with 100% bootstrap support. The red star highlights the branch that separates all the Massoko benthic samples from the rest of the phylogeny (benthic ecomorph individuals are monophyletic in 50% of bootstrap samples). **(B)** Principal Component Analysis of genetic variation within Lake Massoko **(C-E)** Genome-wide pattern of F_{ST} divergence in windows of 15 variants each. Darker color indicates greater density of datapoints. **(C)** Divergence between benthic and littoral ecomorphs within Massoko **(D)** Divergence between combined Massoko and Itamba populations **(E)** Absolute standard scores of Massoko-Itamba divergence (purple) overlaid on divergence between benthic and littoral ecomorphs (green).

Analysis of fine-scale genetic relationships with fineSTRUCTURE (*13*) supports the monophyly of the benthic ecomorph within the littoral, but also suggests that compared with the benthic population, the littoral population has greater coancestry with other *A. calliptera*; in particular with the Mbaka river sample (Fig. S3). Therefore, we tested for evidence of secondary gene flow, as seen in cichlid populations from Cameroonian crater lakes (*14*). Under the null hypothesis of no differential gene flow into Massoko, *A. calliptera* from Mbaka river should share derived alleles equally often with the littoral and with the benthic populations (*15*, *16*). Instead, we found a small excess of shared derived alleles between *A. calliptera* from the Mbaka river and the littoral population, when compared with the benthic population (Patterson's D=1.1%; 4.86 SD from 0% or $P<5.8\times10^{-7}$) (*6*). The proportion of admixture *f* with Mbaka was estimated at 0.9±0.2%. This value is low, at a proportion that is approximately half of the Neanderthal introgression into non-African humans (*15*) and cross-coalescence rate analysis with MSMC (*6*, *17*) indicates an average separation time of both Massoko ecomorphs from other *A. calliptera* samples (including Mbaka river) approximately ten times earlier than the split between the two ecomorphs (Fig. S5). Thus, it is unlikely that a secondary invasion from the neighbouring river systems (Fig. S11B) contributed to the divergence of the ecomorphs.

We estimated individual ancestries for all Massoko and *A. calliptera* specimens with ADMIXTURE (*6*, *18*) (Fig. S4). Focusing on the Massoko samples, 11 of the 31 samples field-assigned as littoral were identified as admixed with admixture fraction >25% from the benthic gene pool. No individuals identified as benthic were estimated to be admixed to the same extent; therefore, recent gene flow may be biased from deep to shallow waters. Ten of the 30 unassigned individuals were also identified as >25% admixed, while the remaining 20 unassigned samples appear to represent sub-adult individuals of both benthic and littoral ecomorphs (Fig. S4A). When additional *A. calliptera* samples were included in ADMIXTURE analysis, a small amount of gene flow into Massoko was apparent with K=2 ancestral populations (Fig. S4C), consistent with the fineSTRUCTURE and Patterson's D results described above. This analysis also suggests similar or even stronger gene flow out of Massoko and into Mbaka river (Fig. S4C).

Islands of speciation

Interestingly, there are no fixed differences between Massoko benthic and littoral ecomorphs. Genome-wide divergence F_{ST} is 0.038, and almost half (47.6%) of the variable sites have zero F_{ST} (Table S7). Above the low background, a genome-wide F_{ST} profile shows clearly demarcated 'islands' of high differentiation (Figs. 2C, 2E). For single sites, the maximum F_{ST} is 13.6 standard deviations (s.d.) above the mean, and 7,543 sites have F_{ST} over 6 s.d. above the mean. By contrast, comparisons of the combined Massoko population and Itamba population revealed a pattern of consistently high F_{ST} across the genome (Fig. 2D) and large variance, making it impossible to detect statistical outliers (Fig. 2E). Similar results were obtained when varying the window size to comprise 15, 50, 100, or 500 variants (Table S7; Figs. S6, S7).

Comparing observed levels of divergence with neutral coalescent simulations (*6*) under a range of possible demographic models revealed that the top 1% of observed F_{ST} values (approximately $F_{ST} \geq 0.25$) are always higher than the corresponding neutral F_{ST} values from simulations, consistent with divergent selection acting on approximately this top 1% of variant sites (Figs. S8, S9).

We identified genomic regions with observed benthic-littoral $F_{ST} \geq 0.25$ (i.e. with F_{ST} above maximum levels seen in neutral simulations) (*6*). For this we used windows of 15 variants each - providing a balance between fine genomic resolution and reducing stochastic variation by averaging over variants. Small gaps arising from brief dips of F_{ST} below the threshold were eliminated by merging regions within 10kb of one another. We found 344 such regions, with total length of 8.1Mb (~1% of the genome). Next, to focus on the more significant outliers, we narrowed the list down to a set of 98 highly diverged regions (HDRs) for further characterization (Table S8) by adding the requirement that at least one 10kb window must have reached $F_{ST} \geq 0.3$. The HDRs vary in length from 4.4kb to 285kb (median 36.1kb), with total length of 5.5Mb.

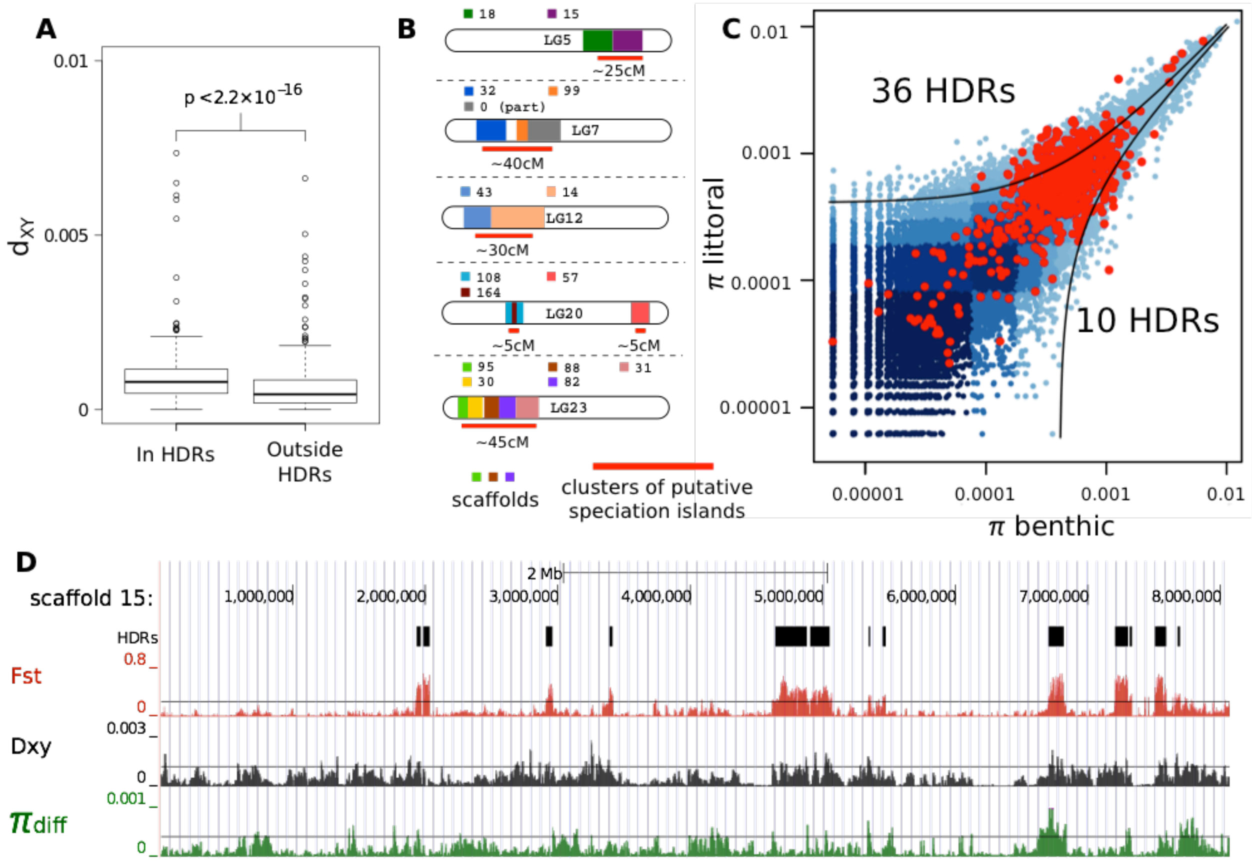


Fig. 3: Islands of speciation between benthic and littoral ecomorphs (A) Elevated d_{XY} in HDRs. **(B)** Clustering of putative speciation islands on five linkage groups. **(C)** Nucleotide diversity (π) within HDRs (red points) and outside HDRs (blue with shading corresponding to density). Each point corresponds to a 10kb window (therefore, there may be multiple points per HDR). Overall 95% of observations lie between the two curves ($y=x\pm4.1\text{x}10^{-4}$). Putative sweeps in the benthic ecomorph are in the top left corner and putative sweeps in the littoral in the bottom right corner. **(D)** Patterns of F_{ST} , d_{XY} , and π_{diff} in a speciation cluster on scaffold 15.

A key prediction of speciation with gene flow models is that loci participating in speciation should have both high relative divergence (F_{ST}) and high absolute sequence divergence (d_{XY}) (*19*, *20*). However, previous studies (examined in ref (*20*)) revealed low d_{XY} and low nucleotide diversity (π) in regions of high F_{ST}. In contrast, we found that d_{XY} in Massoko is significantly higher in HDRs relative to the rest of the genome ($P<2.2\times10^{-16}$, two-tailed Mann-Whitney test; Fig. 3A). In the benthic ecomorph, π in HDRs is not significantly different from the rest of the genome (P=0.34, two-tailed Mann-Whitney test; Fig. S10A), and in the littoral ecomorph π in HDRs is elevated ($P=5.47\times10^{-6}$, two-tailed Mann-Whitney test; Fig. S10B). Individually, 55 HDRs have d_{XY} above the 90th percentile of the genome-wide distribution. The convergence of F_{ST} and d_{XY} measures in regions of normal π suggests that these 55 'islands of speciation' (Table S9) may have been involved in reducing gene-flow in sympatry and thereby directly causing speciation to progress. In contrast, loci involved in continuing local adaptation after the ecomorphs split sufficiently to constitute two largely separate gene-pools (or during a period of allopatry) would be expected to have elevated F_{ST} but not d_{XY} (*20*).

Another key prediction of speciation with gene flow models is that loci causing speciation should be located in relatively few linked clusters within the genome (*2*, *20*, *21*). Instead of a large number of scattered islands, the theory predicts a smaller number of clusters that grow in size due to the 'divergence hitchhiking' process. We tested this prediction using a recently generated linkage map (*22*) and found that at least 27 out of the 55 putative speciation islands are co-localized on five linkage groups (LGs), with 26 of them clustered within their respective LGs (Fig. 3B; Table S9). These potential speciation clusters extended for approximately 25cM on LG5, 40cM on LG7, 30cM on LG12, and 5cM on LG20 and 45cM on LG 23. In total, these regions account for under 7% of the genome, suggesting that divergence hitchhiking may play a role in shaping the observed pattern of genomic differentiation.

Although genomic islands within these clusters are often separated only by a few hundred kb, F_{ST} divergence between HDRs generally drops to background levels (see Fig. 3D), with one exception on scaffold 88 where a broader 'continent' of divergence has formed (Fig. S11).

Further support for sympatric divergence

We next tested whether the HDRs correlated with the signal of gene flow into Lake Massoko, as identified using the sample from the nearby Mbaka river. Compared with the rest of the genome, the HDRs do not have elevated values of Patterson's D (P=0.22, two-tailed Mann-Whitney test; Fig. S12C), nor elevated *f* statistics, which were recently proposed as an means by which one could identify introgressed loci (*6*, *23*) (P=0.08, two-tailed Mann-Whitney test; Fig. S12D). These results suggest that introgression from Mbaka river did not play a major role in generating the HDRs between the benthic and littoral ecomorphs within Lake Massoko (Fig. S12), and strengthen the evidence that the ecomorph divergence has taken place within the lake.

Divergent SNPs associated with mate choice

Many recently diverged taxa, particularly those not geographically isolated, show stronger pre-mating isolation than post-mating isolation (*1*, *24*, *25*). We carried out laboratory experiments to test for reproductive isolation resulting from direct mate choice between the Massoko ecomorphs (*6*) (Movie S2). Fifty Massoko females were given a choice from sixteen males representing the variety of male phenotypes. In parallel, we designed a SNP assay with 117 polymorphic sites representing 44 (HDRs) identified from the first 12 genomes sequenced (Table S10).

We genotyped a reference sample of 18 benthic and 16 littoral males, demonstrating that the SNP assay can reliably separate the ecomorphs along the first principal component (PC1) in PCA (Fig. 4A, top). We then genotyped all females and males participating in the mate-choice experiments (Fig. 4A, bottom) and calculated an average of the PC1 distances between each female and the males she mated with during the experiment, as assayed by microsatellite paternity analysis (*6*). Compared with expectation under random mating (*6*), females had a moderate, but significant ($P=4.3\times10^{-5}$, paired t-test), preference for mating with males more genetically similar to themselves (i.e. closer to them along PC1) (Fig. 4B), demonstrating direct association between HDR variants and mate choice. Assortative mating by genotype was strong among females with positive (littoral) PC1 scores ($P=5.9\times10^{-9}$, paired t-test), while no assortative mating was detected among females with negative (benthic) PC1 scores (Fig. 4B).

Stronger mating discrimination by ancestral populations compared to derived ones has been previously found in *Drosophila* and sticklebacks, possibly because low population density following a founder event favors less choosy individuals (*26*). However, it is also possible that the benthic ecomorph only mates assortatively in the deep water environment; given that our experiments used wide-spectrum lighting characteristic of shallow water. Overall, the moderate assortative mating suggests a role for sexual selection in ecomorph divergence, but does not indicate that it is a primary force causing population-wide divergence.

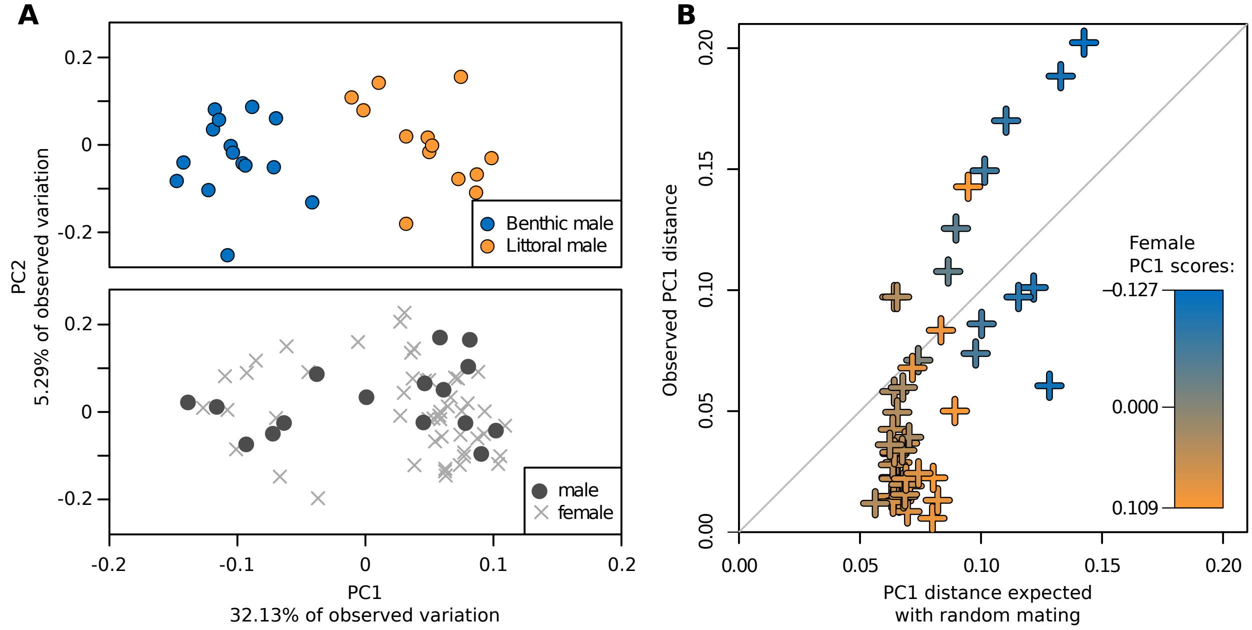


Fig. 4 Mate-choice trials (A) PCA based on 117 genotyped SNPs. **Top:** The first axis of variation (PC1) in PCA reliably separates benthic and littoral males in a reference sample. **Bottom:** PC1 positions of females (N=50) and males (N=16) participating in mate-choice trials. **(B) Results:** Each point compares the average of absolute PC1 distances between a female and: males she mated with (observed PC1 distance) and all males she could have mated with (expected PC1 distance). Points are colored according to the PC1 score of the female. Females below and to the right of the dashed diagonal line on average mate with males more like themselves in terms of PC1 score than would be true if they mated at random.

Signals of adaptation

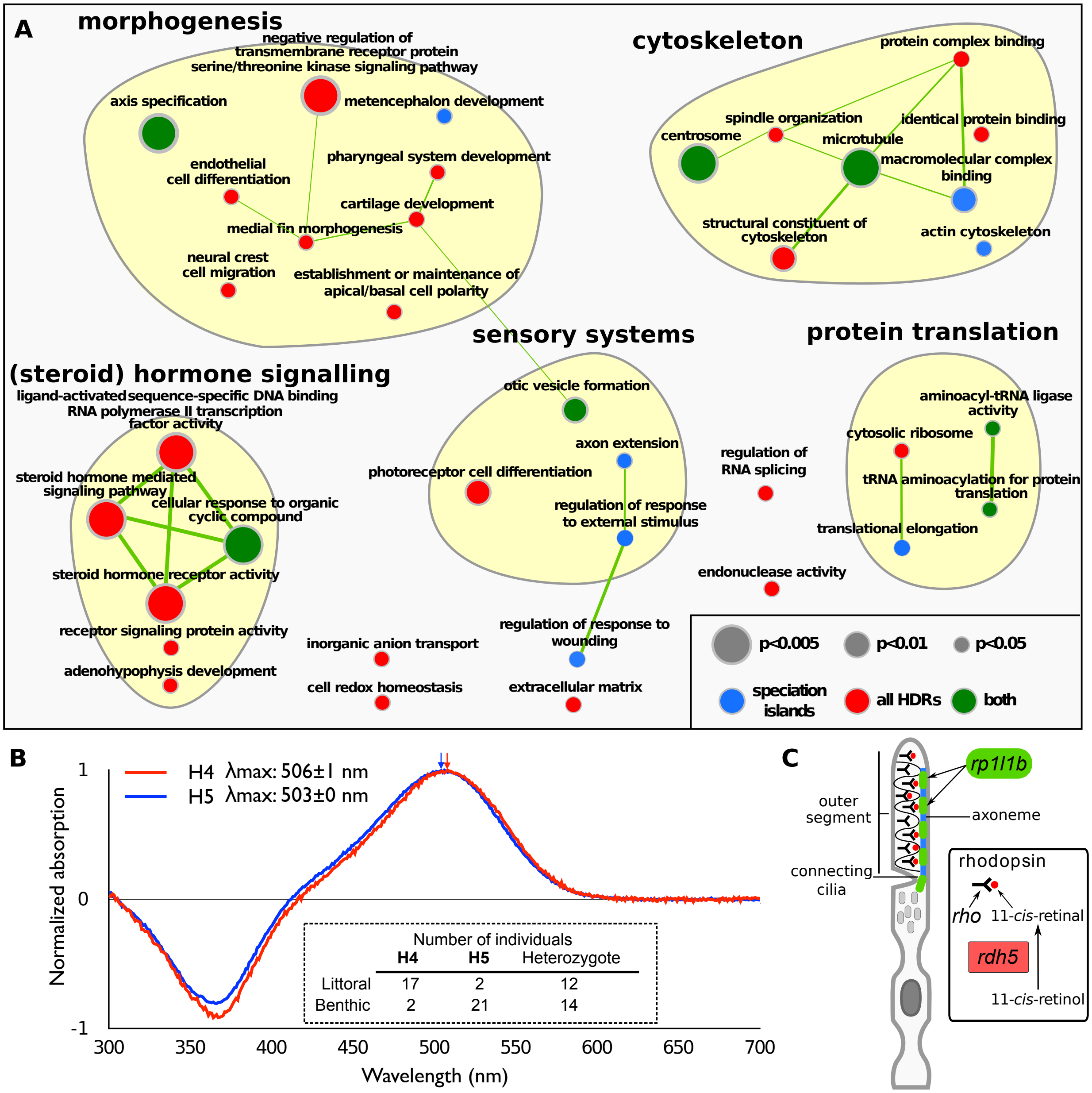
A reduced level of genetic polymorphism in one subpopulation may be indicative of a recent selective sweep. Overall, the magnitude of difference in nucleotide diversity (π) between benthic and littoral ecomorphs (π_{diff}) is significantly higher in the HDRs than in the rest of the genome ($P<2.2\times10^{-16}$, two-tailed Mann-Whitney test; Fig. S13A) (*6*). Individually 46 HDRs have π_{diff} above the 95th percentile of the genome-wide distribution and are likely to have been under recent positive selection in one of the two ecomorphs. There is a significant overlap between HDRs with high d_{XY} (putative 'speciation islands') and HDRs with high π_{diff} (putative recent selective sweeps) - 35 of 55 high d_{XY} islands also have high π_{diff} (Fig. S13B; $P=3\times10^{-5}$, hypergeometric test). On the other hand, the 11 putative sweeps that did not lead to elevated d_{XY} are indicative of adaptation not directly involved in reproductive isolation. Reduced nucleotide diversity in high π_{diff} regions, indicative of selective sweeps, was significantly more prevalent in the benthic ecomorph (36 of 46; $P<1.6\times10^{-4}$, two tailed Binomial test; Figs. 3C, top left; S13C)

(*6*), consistent with the benthic ecomorph being derived and undergoing more extensive adaptation. Nevertheless, there are also a small number of strong outliers suggesting selective sweeps in the littoral ecomorph (Fig. 3C, bottom right).

Functions of adaptation

To explore the function of candidate adaptive genes, we performed Gene Ontology (GO) enrichment analysis (*6*) on three sets: a) genes in candidate 'islands of speciation' ±50kb (enriched terms in Table S11); b) genes in all HDRs ±10kb (Table S12); c) genes in all HDRs ±50kb (Table S13). Combining results of all three analyses in a network (Fig. 5A) connecting GO terms with high overlap (i.e. they share many genes), revealed clear clusters of enriched terms related to: a) morphogenesis (e.g. cartilage and pharyngeal system development, fin morphogenesis), consistent with morphological differentiation; b) sensory systems (e.g. photoreceptor cell differentiation), consistent with previous studies showing the role of cichlid vision in adaptation and speciation (*7*, *27*); and c) (steroid) hormone signalling.

We examined in more detail the functions of candidate genes involved in photoreceptor function (Table S14), and two highly diverged alleles of the rhodopsin (*rho*) gene in Lake Massoko (alleles H4 and H5, separated by four amino acid changes; $F_{ST} = 0.39$; Fig. S14). Blue-shifted rhodopsin absorption spectra are known to play a role in deep-water adaptation (*27*). Therefore, we expressed rhodopsins from H4 and H5 alleles and reconstructed them with 11-*cis*-retinal, measured their absorption spectra (*6*), and demonstrated that the H5 allele, associated with the deep-water benthic ecomorph, has a blue-shifted absorption spectrum (Fig. 5B). The retina-specific retinol dehydrogenase *rdh5* (Table S14) produces 11-*cis*-retinal, the visual pigment binding partner of rhodopsin (*28*), and thus likely has a direct role in dark adaptation. Finally, a mouse ortholog of *rp1l1b* affects photosensitivity and morphogenesis of the outer segment (OS) of rod photoreceptor cells, locating to the axoneme of the OS and of the connecting cilia (*29*) (Fig. 5C). Together, these results suggest divergent selection on *rho*, *rdh5*, and *rp1l1b* may facilitate the adaptation of scotopic (twilight) vision to the darker conditions experienced by the benthic ecomorph.



	Number of individuals		
	H4	H5	Heterozygote
Littoral	17	2	12
Benthic	2	21	14

Fig. 5: Characterizing function of genes in HDRs (A) Enrichment Map for significantly enriched GO terms. The level of overlap between GO enriched terms is indicated by the thickness of the edge between them. The size of the node indicates the best p-value for the term, and the color of the node indicates the gene group for which the term was found significant (i.e. has P<0.05 in candidate 'speciation islands' ±50kb - blue; in all HDRs ±10kb or ±50kb - red; or in both groups - green). Broad functional groupings (morphogenesis, sensory systems…) were derived using automatic clustering followed by manual editing. **(B)** The absorption spectrum of the H5 allele of *rho*, more prevalent in the benthic ecomorph, is shifted towards blue wavelengths. **(C)** The joint roles of *rho*, *rdh5*, and *rp1l1b* in photoreceptor rod cells. *rdh5* produces the chromophore 11-*cis*-retinal that binds *rho*, while *rp1l1b*, located at the axoneme of the outer segment and connecting cilia, also contributes to photosensitivity.

Comparisons to other systems

Overall, our results suggest a pair of incipient species undergoing divergence with gene flow within crater lake Massoko. Their overall level of divergence ($F_{ST} = 0.038$) is low compared with background F_{ST} observed in other recent studies of speciation with gene flow in *Anopheles* mosquitoes (S and M form; $F_{ST} = 0.21$) (*20*), *Ficedula* flycatchers ($F_{ST} = 0.36$) (*30*), and *Heliconius* butterflies ($F_{ST} = 0.18$) (*31*), highlighting that we are looking at an early stage of divergence. The MSMC analysis suggests that median effective divergence occurred within the last 500-1,000 years (~200-350 generations), following separation of lake fish from the Mbaka river population around 10,000 years ago (Fig. S5). However, divergence may have started considerably earlier than these times, masked by subsequent gene flow.

Among populations at similar levels of divergence to Lake Massoko ecomorphs are *Timema* stick insects ($F_{ST} = 0.015$ for adjacent and $F_{ST} = 0.03$ for geographically isolated population pairs), where thousands of regions of moderately elevated divergence were found all across the genome (*32*), and German carrion and Swedish hooded crows ($F_{ST} = 0.017$), that have strongly diverged with fixed differences, but at fewer than five loci (*33*). In Massoko, we observe an intermediate pattern between these two extremes, with a few dozen moderately elevated islands, clustering within the genome indicating close linkage, and no fixed differences. A genome-wide pattern with multiple loci of moderate divergence suggests a genomic architecture similar to the ecological divergence of a sympatric threespine stickleback pair in Paxton Lake, Canada (*34*), and the sympatric divergence of dune-specialist sunflowers, *Helianthus* (*35*).

The ecomorphs of Lake Massoko show clear differences in traits normally associated with adaptive radiation in cichlid fishes, including body shape, pharyngeal jaw morphology, diet, microhabitat preference, retinal pigment sensitivity, male color and mate preference (*3*, *4*, *11*, *12*, *27*). Therefore, our study suggests processes and specific genomic regions for investigations to determine if they are involved in speciation events within the great cichlid radiations of Lakes Malawi, Victoria, and Tanganyika.

References and Notes:

1. J. A. Coyne, H. A. Orr, Speciation (2004).

2. J. L. Feder, S. P. Egan, P. Nosil, The genomics of speciation-with-gene-flow. *Trends in Genetics*. **28**, 342–350 (2012).

3. C. E. Wagner, L. J. Harmon, O. Seehausen, Ecological opportunity and sexual selection together predict adaptive radiation. *Nature*. **487**, 366–369 (2012).

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

5. D. A. Joyce *et al.*, Repeated colonization and hybridization in Lake Malawi cichlids. *Curr. Biol.* **21**, R108–9 (2011).

6. Materials and methods are available as supplementary materials on Science Online.

7. O. Seehausen *et al.*, Speciation through sensory drive in cichlid fish. *Nature*. **455**, 620–626 (2008).

8. M. Delalande, Hydrologie et géochimie isotopique du lac Masoko et de lacs volcaniques de la province active du Rungwe (Sud-Ouest Tanzanie). *http://www.theses.fr* (2008).

9. Barker, Williamson, Gasse, Gibert, Climatic and volcanic forcing revealed in a 50,000-year diatom record from Lake Massoko, Tanzania. *Q Res*. **60**, 9–9 (2003).

10. U. K. Schliewen, D. Tautz, S. Pääbo, Sympatric speciation suggested by monophyly of crater lake cichlids. *Nature*. **368**, 629–632 (1994).

11. M. Barluenga, K. N. Stölting, W. Salzburger, M. Muschick, A. Meyer, Sympatric speciation in Nicaraguan crater lake cichlid fish. *Nature*. **439**, 719–723 (2006).

12. D. Brawand *et al.*, The genomic substrate for adaptive radiation in African cichlid fish. *Nature*. **513**, 375–381 (2014).

13. D. J. Lawson, G. Hellenthal, S. Myers, D. Falush, Inference of population structure using dense haplotype data. *PLoS Genet.* (2012).

14. C. H. Martin *et al.*, Complex histories of repeated gene flow in Cameroon crater lake cichlids cast doubt on one of the clearest examples of sympatric speciation. *Evolution* (2015), doi:10.1111/evo.12674.

15. R. E. Green *et al.*, A draft sequence of the Neandertal genome. *Science*. **328**, 710–722 (2010).

16. E. Y. Durand, N. Patterson, D. Reich, M. Slatkin, Testing for ancient admixture between closely related populations. *Mol. Biol. Evol.* **28**, 2239–2252 (2011).

17. S. Schiffels, R. Durbin, Inferring human population size and separation history from multiple genome sequences. *Nat. Genet.* **46**, 919–925 (2014).

18. D. H. Alexander, J. Novembre, K. Lange, Fast model-based estimation of ancestry in unrelated individuals. *Genes Dev.* **19**, 1655–1664 (2009).

19. M. A. F. Noor, S. M. Bennett, Islands of speciation or mirages in the desert? Examining the role of restricted recombination in maintaining species. *Heredity*. **103**, 439–444 (2009).

20. T. E. Cruickshank, M. W. Hahn, Reanalysis suggests that genomic islands of speciation are due to reduced diversity, not reduced gene flow. *Mol Ecol*. **23**, 3133–3157 (2014).

21. S. Via, Divergence hitchhiking and the spread of genomic isolation during ecological speciation-with-gene-flow. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*. **367**, 451–460 (2012).

22. C. T. O Quin, A. C. Drilea, M. A. Conte, T. D. Kocher, Mapping of pigmentation QTL on an anchored genome assembly of the cichlid fish, *Metriaclima zebra*. *BMC Genomics*. **14**, 287 (2013).

23. S. H. Martin, J. W. Davey, C. D. Jiggins, Evaluating the use of ABBA-BABA statistics to locate introgressed loci. *Mol. Biol. Evol.* **32**, 244–257 (2015).

24. T. D. Price, M. M. Bouvier, The evolution of F1 postzygotic incompatibilities in birds. *Evolution*. **56**, 2083–2089 (2002).

25. R. B. Stelkens, K. A. Young, O. Seehausen, The accumulation of reproductive incompatibilities in African cichlid fish. *Evolution*. **64**, 617–633 (2010).

26. K. Y. Kaneshiro, Sexual Isolation, Speciation and the Direction of Evolution. *Evolution*. **34**, 437 (1980).

27. T. Sugawara *et al.*, Parallelism of amino acid changes at the RH1 affecting spectral sensitivity among deep-water cichlids from Lakes Tanganyika and Malawi. *Proc. Natl. Acad. Sci. U.S.A.* **102**, 5448–5453 (2005).

28. G. Duester, Families of retinoid dehydrogenases regulating vitamin A function: production of visual pigment and retinoic acid. *Eur. J. Biochem.* **267**, 4315–4324 (2000).

29. T. Yamashita *et al.*, Essential and synergistic roles of RP1 and RP1L1 in rod photoreceptor axoneme and retinitis pigmentosa. *J. Neurosci.* **29**, 9748–9760 (2009).

30. H. Ellegren *et al.*, The genomic landscape of species divergence in *Ficedula* flycatchers. *Nature*. **491**, 756–760 (2012).

31. N. J. Nadeau *et al.*, Genomic islands of divergence in hybridizing *Heliconius* butterflies identified by large-scale targeted sequencing. *Philos. Trans. R. Soc. Lond., B, Biol. Sci.*

367, 343–353 (2012).

32. V. Soria-Carrasco *et al.*, Stick insect genomes reveal natural selection's role in parallel speciation. *Science*. **344**, 738–742 (2014).

33. J. W. Poelstra *et al.*, The genomic landscape underlying phenotypic integrity in the face of gene flow in crows. *Science*. **344**, 1410–1414 (2014).

34. M. E. Arnegard *et al.*, Genetics of ecological divergence during speciation. *Nature*. **511**, 307–311 (2014).

35. R. L. Andrew, L. H. Rieseberg, Divergence is focused on few genomic regions early in speciation: incipient speciation of sunflower ecotypes. *Evolution*. **67**, 2468–2482 (2013).

Acknowledgments: The work was funded by Royal Society-Leverhulme Trust Africa Awards AA100023 and AA130107 (M.J.G., B.P.N. and G.F.T.), a Wellcome Trust PhD studentship grant 097677/Z/11/Z (M.M.), Wellcome Trust grant WT098051 (S.S. and R.D.), Wellcome Trust and Cancer Research UK core support and a Wellcome Trust Senior Investigator Award (E.A.M.), a Leverhulme Trust Research Fellowship RF-2014-686 (M.J.G.), a University of Bristol Research Committee award (M.G.), a Bangor University Anniversary PhD studentship (to A.M.T.) and a Fisheries Society of the British Isles award (G.F.T.). Raw sequencing reads are in the SRA nucleotide archive: RAD sequencing (BioProject: PRJNA286304; accessions SAMN03768857 to SAMN03768912) and whole genome sequencing (BioProject PRJEB1254: sample accessions listed in Table S16). The RAD based phylogeny and alignments have been deposited in TreeBase (TB2:S18241). Whole genome variant calls in the VCF format, phylogenetic trees, and primer sequences for Sequenom genotyping are available from the Dryad Digital Repository (http://dx.doi.org/10.5061/dryad.770mc). RD declares his interests as a founder and non-executive director of Congenica Ltd., that he owns stock in Illumina from previous consulting, and is a scientific advisory board member of Dovetail Inc. We thank R. Schley for generating pharyngeal jaw data; S. Mzighani, J. Kihedu and staff of the Tanzanian Fisheries Research Institute for logistical support; A. Smith, H. Sungani, A. Shechonge, P. Parsons, J. Swanstrom, G. Cooke and J. Bridle for contributions to sampling and aquarium maintenance, the Sanger Institute sequencing core for DNA sequencing and Dr. H. Imai (Kyoto University) for the use of spectrometer in his laboratory.

Supplementary Materials:

Materials and Methods
Figures S1-S16
Tables S1-S16
References (*36-67*)
Movies S1,S2

Science
AAAS

Supplementary Materials for

Genomic Islands of Speciation Separate Cichlid Ecomorphs in an East African Crater Lake

Milan Malinsky, Richard Challis, Alexandra M. Tyers, Stephan Schiffels, Yohey Terai, Benjamin P. Ngatunga, Eric A. Miska, Richard Durbin, Martin J. Genner, George F. Turner

correspondence to: george.turner@bangor.ac.uk, M.Genner@bristol.ac.uk

This PDF file includes:

Materials and Methods
Figs. S1 to S16
Tables S1 to S16
Captions for Movies S1 to S2

Other Supplementary Materials for this manuscript includes the following:

Movies S1 to S2

Materials and Methods

1. Field sampling and eco-morphological analysis

Field sampling for genetic, morphological and stable isotope samples:
Astatotilapia samples from Lake Massoko were collected on 17th July 2011, and from 19th to 25th November 2011. Fish were collected using fixed gill nets and SCUBA. On being brought to the surface, fish were given an overdose of anesthetic (MS-222). From each fish we collected a genetic sample (fin clip) that was stored in ethanol, and cut a fillet of the flank for stable isotope analyses that was sun-dried and stored with desiccant. Samples of potential food sources were also collected and dried, including epilithic algae, sponges and bivalves. Whole fish were preserved in formalin (~4%). Genetic samples (fin clips) of outgroup *Astatotilapia calliptera* were collected opportunistically between 2009 and 2014 (Fig. S2; Tables S2, S4).

Field sampling for assessment of ecomorph frequency with depth:
Astatotilapia samples from Lake Massoko were collected from 28th July to 7th August 2014, 10th to 15th December 2014, and 5th - 24th August 2015. Fish were collected using fixed gill nets, angling and SCUBA. Depth was assessed using a plumbline, surface depth meter, and dive gauges. Fish were photographed on collection. All adult males >65mm Standard Length were assigned to an ecomorph on the basis of the color of body and fins, and gross morphology. Depth was recorded as bottom depth, which means that fish caught in the water column may sometimes be included. This, along with drift of passive fishing gears may perhaps have led to an over-representation of shallow water fish in deeper water records.

Live fish collection:
Live fish were collected in November 2011 by a team of professional aquarium fish collectors, under our supervision, for shipment to UK for mate choice trials. These were collected from depths of >20m or < 5m to ensure good representation of both ecomorphs. Deep-water fish were decompressed overnight in keep-nets at depths of 5-10m.

Morphological analysis:
To quantify differences in body morphology among individuals we used a landmark-based morphometric approach that captured the variation in the geometric relationships among defined points. All individuals were photographed in a standard orientation with the head pointing left. Images were calibrated to scale and 22 landmarks (Fig. S15) were marked using `tpsDig2.17` (*36*). Coordinates were aligned using Procrustes analysis in `MorphoJ 1.06d` (*37*), enabling the generation of Principal Component scores along the primary axis of morphological variation. Lower pharyngeal jaws were removed from the fish, and weighed to the nearest 0.1mg using a METTLER TOLEDO AB54-S balance. A photograph was then taken of the jaw and a distance scale using a ZEISS AXIOSTAR light microscope with a fitted NIKON D70 camera at 3.5x magnification.

The width of the posterior three teeth immediately to the right side of the suture were measured using `tpsDig2.17` to the nearest 0.01mm.

Stable isotopes measurements:

Carbon-13 and Nitrogen-15 Elemental Analysis was conducted using Isotope Ratio Mass Spectrometry at IsoAnalytical, Crewe UK. The technique used for stable isotope analysis was Elemental Analysis - Isotope Ratio Mass Spectrometry (EA-IRMS) using a Europa Scientific 20-20 IRMS. Carbon results were recorded relative to the Vienna Peedee Belemnite scale (V-PDB). The reference material used for $\delta^{13}C$ and $\delta^{15}N$ analysis was IA-R042 (NBS-1577B, powdered bovine liver, $\delta^{13}C_{V\text{-}PDB} = -21.60$ ‰, $\delta^{15}N_{AIR} = 7.65$ ‰). Additionally, samples were run against multiple reference materials for confirming accuracy of the results:

1. IA-R032 = powdered bovine liver, the maximum deviation observed from the accepted reference values was 0.14 ‰ for $\delta^{15}N$, and 0.14 ‰ for $\delta^{13}C$.
2. IA-R045 / IA-R005 = a mixture of ammonium sulphate and beet sugar, the maximum deviation observed from the accepted reference values was 0.15 ‰ for $\delta^{15}N$, and 0.17 ‰ for $\delta^{13}C$.
3. IA-R046 / IA-R006 = a mixture of ammonium sulphate and beet sugar, the maximum deviation observed from the accepted reference values was 0.22 ‰ for $\delta^{15}N$, and 0.14 ‰ for $\delta^{13}C$.

Every 5th sample was repeated as a control. The maximum difference observed between replicates was 0.16 ‰ for $\delta^{15}N$, and 0.23 ‰ for $\delta^{13}C$. In total we analyzed 46 individuals of the focal Lake Massoko *Astatotilapia* ecomorph pair, and 10 samples of potential food sources (Fig. S16).

2. RAD-seq data processing and analysis:

DNA extraction and sequencing:

DNA was extracted from ethanol-preserved fin tissue from 56 wild caught fish using a standard CTAB-Chloroform extraction method including an RNAase treatment step. This was sent to Floragenex (http://www.floragenex.com/) for library preparation using the Sbf1 enzyme and sequencing on an Illumina HiSeq2000 platform, providing 100bp single end reads. The samples were sequenced in two rounds. In the first round sequencing was 28 samples per lane, but 43 individuals obtained less than 1M reads each. In a second round 41 of these 43 individuals were reprepped, and sequenced at 20 and 21 samples per lane. Raw data have been deposited at the NCBI Sequence Read Archive under BioProject PRJNA286304 (Accessions SAMN03768857 to SAMN03768912).

Variant calling and filtering:

Samples with fewer than 300000 reads (approximately 20X coverage per tag) were removed. Raw reads for the remaining 42 samples were de-multiplexed and adaptor trimmed leaving 89 base reads for use in reference guided RAD tag analysis.

Reads were aligned to the Mbaka River *Astatotilapia calliptera* consensus sequence (see Whole genome data processing and analysis) using `bwa-mem v.0.7.12` (*38*). An average of 96.2% (±0.3%) of reads mapped to the reference and these mapped reads were filtered to remove reads with terminal alignments and reads that were not uniquely mappable leaving an average of 90.3% (±1.1%) of the original reads in the filtered read set. SNPs were called using the `stacks` (*39*) `ref_map.pl` pipeline with a minimum stack depth (`-m`) of 5. The full dataset was filtered to remove SNPs that had been called in less than 75% of samples and the resulting matrix contained 7,906 SNPs and was 82.3% complete.

Phylogenetic trees and constraint tests:

Phylogenetic model testing using `ModelGenerator v.0.85` (*40*) supported the use of the GTR + Γ model of sequence evolution with an estimated transition/transversion ratio of 2.65. A maximum likelihood (ML) phylogeny was produced using `RAxML v.8.0.22` (*41*) using the `GTRGAMMA` model. Support for the ML tree topology was inferred using 100 rapid bootstrap samples (*42*). The phylogeny was rooted on *A. tweddlei* and has been deposited in TreeBase (accession: TB2:S18241).

The hypothesis of monophyly of each of the crater lakes (Table S15) was tested by generating an ML phylogeny with the lake constrained to be monophyletic (RAxML option `-g MASSOKO_MONOPHYLY`, etc.). Each of the constrained topologies was compared with the unconstrained using the Shimodaira-Hasegawa (SH) test (*43*) in RAxML using the command:

```
raxml -T 4 -f h -t UNCONSTRAINED_TREE -z CONSTRAINT_TREES -s MATRIX.phy -m
GTRGAMMA -n TEST
```

3. Whole genome data processing and analysis:

DNA extraction and sequencing:

DNA was extracted from fin clips using PureLink® Genomic DNA extraction kit (Life Technologies). Genomic libraries for paired-end sequencing on the Illumina HiSeq 2000 machine were prepared according to Illumina TruSeq HT protocol to obtain paired-end reads with mean insert size of 300-500bp. As detailed in Table S6, we used either Illumina HiSeq v3 chemistry (generating 100bp paired-end reads) or Illumina HiSeq v4 reagents (125bp paired-end reads). Low coverage (~6x) samples with v4 reagents were multiplexed 12 per lane. High coverage (~15x) v4 samples were multiplexed four per lane. For high coverage (~15x) v3 samples, a multiplexed library with 8 samples was sequenced over three lanes. Raw data have been deposited at the NCBI Sequence Read Archive under BioProject PRJEB1254; individual sample accessions are listed in Table S16.

Alignment:

All reads were aligned to *Metriaclima zebra* reference genome (*12*) using the `bwa-mem v.0.7.10` algorithm (*38*) using default options. For each sample, 96-98% of reads could be aligned to the reference. Duplicate reads were marked on both per-lane and per sample basis using the `MarkDuplicates` tool from the `Picard` software package with default options (http://broadinstitute.github.io/picard) and local realignment around indels performed on both per lane and per sample basis using the `IndelRealigner` tool from the GATK v.3.3.0 software package (*44*).

Variant calling, filtering, and genotype refinement:

Briefly, SNP and short indel variants against the *M. zebra* reference were called independently using GATK v3.3.0 haplotype caller (*45*) and samtools/bcftools v.1.1 (*46*). Variant filtering was then performed on each set of variants separately using hard filters based on overall depth, overall quality score, strand/mapping bias, and inbreeding coefficient (see below). Multiallelic sites were excluded. After filtering, we selected consensus sites (i.e. we performed intersection of GATK and samtools sites). At a particular locus, if the GATK and samtools alleles differed, we kept the GATK allele. Finally, we used genotype likelihoods output by GATK at consensus sites to perform genotype refinement, imputation, and phasing in BEAGLE v.4.0 (*47*). Except where specifically indicated, indels were excluded from analyses using `vcftools v0.1.12b` option `--remove-indels.`

The particular commands/parameters used were:

samtools calling (multisample):

```
samtools   mpileup   -t   DP,DPR,INFO/DPR   -C50   -pm2   -F0.2   -ugf   REFERENCE.fa
SAMPLE1.bam   SAMPLE2.bam   …   |   bcftools   call   -vmO   z   -f   GQ   -o
samtools_VARIANTS.vcf.gz
```

GATK haplotype caller (per sample), later combined using GATK's `GenotypeGVCFs` tool:

```
java   -jar   GenomeAnalysisTK.jar   -T   HaplotypeCaller   -R   REFERENCE.fa   --
emitRefConfidence  GVCF  --variant_index_type  LINEAR  --variant_index_parameter
128000 -I SAMPLEn.bam -o GATK_SAMPLEn.g.vcf
```

Hard filters applied to both datasets:

```
Minimal inbreeding coefficient: -0.05
Minimum overall read depth: 600
Maximum overall read depth: 1700 (except for mtDNA: scaffolds 747,2036)
```

Hard filters applied to the GATK dataset:

```
Maximum phred-scaled p-value using Fisher's exact test to detect strand bias:
20 (except for mtDNA: scaffolds 747,2036)
Minimum accepted variant quality score: 300
```

Hard filters applied to the samtools dataset:

```
Minimum p-value for Mann-Whitney U test of Mapping Quality vs. Strand Bias:
0.0001
(except for mtDNA - scaffolds 747,2036)
Minimum accepted variant quality score: 30
```

The consensus GATK and samtools call set was obtained using the bcftools `isec` tool:

```
bcftools     isec     -c     indels     -O     z     GATK_filtered_calls.vcf.gz
samtools_filtered_calls.vcf.gz -p GATK_samtools_intersect/
```

BEAGLE genotype refinement (per scaffold):

```
java   -jar   beagle.r1398.jar   gl=GATK_samtools_consensus.vcf.gz   phase-its=8
impute-its=8 out=beagle_GATK_sam_consensus
```

Whole genome phylogenetic trees:

Consensus genome sequences were generated using the `bcftools v1.2 consensus` tool. For each sample, the sequence of one haplotype was selected (as assigned by beagle haplotype phasing - see above) by using the `-haplotype=1` option in `bcftools`. All scaffolds except the mtDNA sequence (scaffolds 747, 2036) were concatenated into a single sequence and phylogenetic trees then inferred using `RAxML v7.7.8` (*41*) under the GTRGAMMA model (General Time Reversible model of nucleotide substitution with the Γ model of rate heterogeneity). The maximum likelihood tree was obtained as the best out of five alternative runs on distinct starting maximum parsimony trees (using the `-N 5` option). Sixty six bootstrap replicates were obtained using RAxML's rapid bootstrapping algorithm (*42*), giving a reasonable indication of bootstrap support for the maximum likelihood tree (obtaining the 66 replicates required ~7,647 hours of CPU time). Bipartition bootstrap support was drawn on the maximum likelihood tree using RAxML `-f b` option.

Principal Component Analysis:

SNP variants with minor allele frequency >= 0.05 were selected using `vcftools v0.1.12b` options `--maf 0.05` and exported in PLINK format (*48*). The variants were LD-pruned to obtain a set of variants in approximate linkage equilibrium (unlinked sites) using the `--indep-pairwise 50 5 0.2` option in `PLINK v1.0.7`. Principal Component Analysis on the resulting set of variants was performed using the `smartpca` program from the `eigensoft v5.0.1` software package (*49*) with default parameters.

ADMIXTURE ancestry estimation:
All SNP variants were exported in PLINK format (*48*) and LD-pruned to obtain a set of variants in approximate linkage equilibrium (unlinked sites) using the `--indep-pairwise 50 5 0.2` option in `PLINK v1.0.7`. The `ADMIXTURE v1.23` program (*18*) was then run with default parameters. The postulated number of ancestral populations `K` was set to 1, 2, 3, 4, 5, and 6. From a statistical standpoint, the authors of the software suggest choosing the value of `K` with the lowest cross-validation error. We performed 10-fold cross-validation (`--cv=10`) and found the lowest cross-validation error is with `K=1` (when using Massoko data only; Fig. S4A). ADMIXTURE cross-validation implying `K=1` is a common phenomenon when population differentiation is subtle, but meaningful results can still be obtained with higher values of `K` - see for example the application of ADMIXTURE to HGDP human European data in (*13*) and Figure S12 therein.

Chromopainter and fineSTRUCTURE:
Singleton SNPs were excluded using `bcftools-1.1 -c 2:minor` option, before exporting the remaining variants in PLINK format (*48*). The `chromopainter v0.0.4` software (*13*) was then run for 150 largest genomic scaffolds. Briefly, we created a uniform recombination map using the `makeuniformrecfile.pl` script, then estimated the effective population size (N_e) for a subsample of 20 individuals using the `chromopainter` inbuilt expectation-maximization procedure (*13*), averaged over the 20 N_e values using the provided `neaverage.pl` script. Estimated Ne values ranged from 1,046 to 6,015 (mean 3914, sd. 990). The `chromopainter` program was then run for each scaffold independently, with the `-a 0 0` option to run all individuals against all others. Results for individual scaffolds were combined using the `chromocombine` tool before running `fineSTRUCTURE v0.0.5` with 1,000,000 burn in iterations, and 200,000 sample iterations, recording a sample every 1,000 iterations (options `-x 1000000 -y 200000 -z 1000`). Finally, the sample relationship tree was built with `fineSTRUCTURE` using the `-m T` option and 20,000 iterations.

Patterson's D (ABBA-BABA) and related statistics:
To test for possible gene-flow between surrounding rivers and Massoko, we calculated the ABBA-BABA statistic (*15*, *16*). The ABBA-BABA test (also known as 'D statistic' or "Patterson's D") tests for evidence of introgression in the form of an excess of shared derived alleles between one of two populations and an outgroup. Formally, we calculated D(benthic, littoral, Mbaka river, *P. nyererei*) using equation S15.2 of Green *et al.* (*15*), allowing us to use allele-frequency information from all benthic and littoral individuals. We also estimated *f*, the admixture fraction following Green *et al.* equation S18.5, and calculated the standard error for both estimates by a weighted block jackknife, using blocks of 5,000 informative variants (i.e. variants with ABBA or BABA patterns).

Finally, we calculated a version of the f_d statistic designed by Martin *et al.* specifically to detect introgressed loci (*23*), equation 6. The f_d statistic is distributed on the interval (-Infinity,1]. We also define a closely related statistic which we call f_{dM}. Compared with the f_d statistic, f_{dM} has the advantages that it is bounded on a the interval [-1,1], and under the null hypothesis of no introgression is symmetrically distributed around zero.

Following the notation from Martin *et al.* (*23*), we consider three populations and an outgroup with the relationship $(((P_1, P_2), P_3), O)$. Then:

$S(P_1;P_2;P_3;O) = \mathbf{\Sigma_i}((1-p_{i1})*p_{i2}*p_{i3}*(1-p_{i4}))-\mathbf{\Sigma_i}(p_{i1}*(1-p_{i2})*p_{i3}*(1-p_{i4}))$

where p_{ij} is the frequency of the derived allele at site i in population j.

f_{dM} is then defined as follows:

- if $p_{i2} >= p_{i1}$ then $f_{dM}=f_d=S(P_1;P_2;P_3;O)/S(P_1,P_D,P_D,O)$
- if $p_{i1} < p_{i2}$ then $f_{dM}=S(P_1;P_2;P_3;O)/-S(P_D,P_2,P_D,O)$

where P_D is the population (either P_1 or P_3) that has the higher frequency of the derived allele. For a detailed discussion of the f_d statistic see Martin *et al.* (*23*).

The D and *f* statistics were calculated genome-wide and D and f_{dM} also in non-overlapping windows of 50 informative variants each.

To obtain ancestral allele information, we generated a whole genome alignment between *M. zebra* and *P. nyererei* (cichlid from Lake Victoria, several million years diverged) genome assemblies (*12*). Briefly, alignments were generated using `lastz v1.02`, with the following parameters:

```
B=2  C=0  E=150  H=0  K=4500  L=3000  M=254  O=600  Q=human_chimp.v2.q  T=2
Y=15000
```

followed by using Jim Kent's `axtChain` tool with `-minScore=5000` and additional tools with default parameters in order to obtain a contiguous alignment on *M. zebra* genomic coordinates following the UCSC whole-genome alignment paradigm (http://genomewiki.ucsc.edu/index.php/Whole_genome_alignment_howto). Finally, indels were removed from the alignment and ancestral allele information for SNPs filled into the VCF file using our custom C++ program `evo` with the `aa-seq` and `aa-fill` options (available from `https://github.com/millanek/evo`).

MSMC cross-coalescence analysis:

Because results of this analysis rely, in part, on detecting the density of heterozygous sites, we restricted this analysis to high coverage (~15X) samples. Genomic regions on which short reads cannot be uniquely mapped were masked out by a) excluding genomic regions where mapped depth was higher than 35X (more than twice the average genome coverage); b) using Heng Li's SNPable tool (http://lh3lh3.users.sourceforge.net/snpable.shtml). The SNPable tool divides the reference genome into overlapping *k*-mers (sequences of length *k* – we used `k=50`), and then the extracted *k*-mers are aligned back to the genome (we used `bwa aln -R`

`1000000 -O 3 -E 3`). Then we only kept regions where the majority of overlapping 50-mers were mapped back uniquely and without 1-difference.

Running `MSMC` without the `--fixedRecombination` parameter for 100 iterations indicated that the `--rhoOverMu` parameter is approximately 2 (the parameter name is misleading, as it refers to the ratio r/μ, where r is the per generation recombination rate per base pair (bp) and μ the per generation mutation rate per bp. This value was used for all following MSMC runs (`msmc --rhoOverMu=2 --fixedRecombination`).

Each run of the cross-coalescence analysis used four haplotypes, two from each ecomorph for the benthic-littoral split, and two from Mbaka and two from Massoko for the Massoko-Mbaka split.

Since `MSMC` relies on long-range haplotype phasing, we re-phased the data using the `shapeit v2.r790` haplotype phasing method (*50*) including the use of phase-informative reads (*51*). Because of the need for long-range phase information, we restricted the analysis to 50 largest genomic scaffolds, comprising ~390Mb of sequence.

Neutral coalescent simulations:

We used the coalescent simulator `ms` (*52*) to simulate the divergence of two subpopulations, sampling 74 chromosomes from the first population corresponding to 37 Massoko benthic samples and 64 chromosomes from the second population corresponding to 32 Massoko littoral samples (`-I 2 74 64` parameter). The simulations were performed under a range of models and demographic scenarios, as described in the main text. Migration rate for the Isolation with migration (IWM) model was included directly in the `-I` parameter (e.g. `-I 2 74 64 5`) and for the Isolation after migration model was adjusted using the `-eM` option.

For each model/scenario: a) the between-population split time (`-ej` parameter) was adjusted to match the overall observed benthic littoral F_{ST} of 3.89%; b) we simulated 500,000 independent samples, each sample with one segregating site (effectively simulating 500 thousand unlinked loci). Therefore, the basic command line for the IWM model looked as follows:

```
ms 138 500000 -s 1 -I 2 74 64 M -ej splitT 1 2
```

where `M` is the migration parameter, and `splitT` stands for the split time.

Calculating F_{ST} and defining HDRs:

F_{ST} was calculated both for simulations and for the cichlid data using our custom C++ program `evo` (available from `https://github.com/millanek/evo`). The `fst --ms` option was used for simulations and the `fst --vcf` option for cichlid data. Both SNP variants and indels were used for these analyses. Our F_{ST} calculation implements the Hudson estimator, as defined by Bhatia, Patterson *et al.* (*53*) in equation 10, using 'ratio of averages' to combine estimates of F_{ST} across multiple variants, as recommended in their manuscript.

For defining HDRs, we used windows of 15 variants each, which we found to provide good balance between fine genomic resolution and reducing stochastic variation by averaging over variants. Nevertheless, we found some cases where F_{ST} between neighboring regions dipped briefly below the threshold, which we believe to be in most cases due to remaining stochastic variation. The length (the extent) of HDRs was defined by merging windows with $F_{ST} >= 0.25$ that were next to each other or within 10,000bp of one another using `bedtools v2.16.2` (*54*): `mergeBed -d 10000 -i windows_fst_above0.25.bed`. F_{ST} was also calculated in 10kb windows and each HDR must contain at least one window with $F_{ST} >= 0.3$, as described in the main text.

Characterization of HDRs in terms of d_{XY}, π, and π_{diff}:

Both d_{XY} and nucleotide diversity (π) were calculated for 10kb windows. The d_{XY} statistic was calculated as defined by Wakeley (*55*) in equation 3. Both calculations are implemented in our custom C++ program `evo` (available from `https://github.com/millanek/evo`), and were obtained by using the `fst --vcf` option.

Average nucleotide diversity in each window was calculated separately for the benthic (π_B) and littoral (π_L) ecomorphs and π_{diff} was then calculated as the absolute value of the difference between π_B and π_L; i.e. $\pi_{diff} = |\pi_B - \pi_L|$. The 'direction' of the 'sweep' is in the morph with lower π; i.e. if $\pi_B < \pi_L$ then the potential 'sweep' was inferred to be in the benthic morph.

Gene Ontology enrichment analysis:

We used the `V1` gene annotations generated at the Broad Institute as a part of the cichlid genome project (*12*), including assignment of orthologs between the *M. zebra* genome and zebrafish (*Danio rerio*). Zebrafish has the most extensive functional gene annotation of any fish species, providing a basis for Gene Ontology (GO) (*56*) term enrichment analysis. Genome-wide, 13,230 (61.3%) of *M. zebra* genes had an assigned zebrafish ortholog, mapping to 11,810 unique zebrafish genes.

Gene Ontology (GO) enrichment for genes found within HDRs was calculated in `R` using the `topGO` package (*57*) from the `Bioconductor` project (*58*). The GO hierarchical structure was obtained from the `GO.db v3.1.2` annotation and linking zebrafish gene identifiers to GO terms was accomplished using the `org.Dr.eg.db v3.1.2` annotation package. Genome-wide, approximately 7,000 genes had a GO annotation that could be used by `topGO`, the exact number depending on the GO category being assessed. The `nodeSize` parameter was set to 10 to remove GO terms which have less than 10 annotated genes, as suggested in the `topGO` manual.

There is often an overlap between gene-sets annotated with different GO terms, in part because the terms are related to each other in a hierarchical structure (*56*). Therefore, we used the Enrichment Map (*59*) app for Cytoscape (http://www.cytoscape.org) to organize all the significantly enriched terms into networks where terms are connected if they have a high overlap, i.e. if they share many genes. Broader functional groupings (morphology, sensory systems etc.) were initially derived using clusterMaker (*60*) and WordCloud (*61*), followed by manual editing.

4. Mate choice experiments:

Experimental setup, and aquarium work:
A single 4m long tank with a gravel/sand substrate was divided into eight sections by ‘partial partition’ grids (*62*). Each section contained a terracotta plant pot (to function as territorial focal points for the males) and a selection of plastic plants. Water was filtered and heated (to ~26 C) externally and the tank lit from above by white and UV enhanced fluorescent tube lamps. Fish were fed daily with algae flake and 2-3 times weekly with frozen bloodworm.

Two female mate choice trials were carried out using two different sets of eight males and a total of 50 females. All fish were wild caught and shipped to the UK in December 2011. Trial 1 ran from the beginning of November 2012 to the end of January 2013 (3 months) and trial 2 started at the beginning of February 2013, ending in June 2013 (4.5 months). Each set/trial comprised 3-4 large littoral, 1-2 large benthic and 3-4 'small' males. Forty-five of the 50 females produced broods in both trials. All of the larger littoral and benthic males within each set were of a comparable size and unable to fit through the partial partition grids. The large males were placed in every-other section, leaving the territories in-between available to the small males which, being of a similar size to some of the larger females, were also able to move freely between sections. Before introduction of the females to the experimental tank, males were left until the smaller ones had settled into the 'empty' territories between the bigger males.

As with other haplochromine cichlid fish, *Astatotilapia* are maternal mouth-brooders, egg are picked up by the female during spawning and protected in the buccal cavity during development before release as free-swimming young approximately three weeks later. Females were removed from the experimental tank after spawning and isolated in small tanks on a recirculating system during the brooding phase. After the first trial, offspring were gently removed from the females mouths after 10 days and euthanized by anesthetic (clove oil) overdose. Females were kept in their individual tanks to allow for rest and recovery before the second trial. Once all females had spawned in the first trial, the males were changed. Allele diversity at the chosen microsatellite loci (Ppun5, 7 & 21) (*63*) was sufficiently high to allow for the identification of all individual females by their microsatellite profile, it was therefore possible to return all females to the experimental tank at the same time for the second trial and re-identify individuals later during the second round of paternity testing. After spawning in the second trial, females were again isolated, but left to brood to term. Five offspring from each brood were euthanized for paternity testing.

Paternity testing
475 offspring from 95 broods (five per brood/trial), produced over the two replicates were genotyped for paternity analysis (250 from 50 broods in trial 1; 225 from 45 broods

in trial 2). Tissue was taken from ethanol preserved fry samples and DNA obtained by salt extraction. DNA samples from offspring, mothers, and all potential fathers were used for assigning paternity by allele sizing after PCR multiplex (Qiagen multiplex kit) of three microsatellite markers (Ppun5, 7 & 21) (*63*). Genotyping of the amplified samples was carried out on an Applied Biosystems (ABI) 3130xl genetic analyzer using LIZ 500(-250) (ABI) size standard. The genotype of each individual (males, females, offspring) were determined by manual scoring of alleles in `Peak Scanner v2`. 447 (94%) of the genotyped offspring were successfully assigned to an individual male. Due to allele sharing among males used in trial 2, 23 offspring could not be assigned unambiguously. Seven offspring could not be assigned due to problems with amplification or disagreement between microsatellite loci (possible cross-contamination). Overall, 8-10 offspring from each female that produced more than one brood, were unambiguously assigned to father.

Forty-six of the 50 females spawned with more than one male during the course of the experiment and some females were found to have spawned with up to four males in total (44% of individual broods were sired by more than one male).

Data analysis

We designed a Sequenom MassARRAY SNP genotyping assay (*64*) for 117 SNPs (Table S10), over four Sequenom plates. The assay was performed by the Wellcome Trust Sanger Institute core genotyping team. Analysis of the SNP data was performed in R using the `Bioconductor` (*58*) package `SNPRelate` (*65*) to account for linkage disequilibrium (LD) between the SNPs. SNPs were filtered using a recursive sliding window approach (`snpgdsLDpruning`) with an LD threshold of 0.2. Principal components analysis of the filtered dataset was used to obtain a score for each of the individuals used in the mate choice experiments.

We calculated expected distance between female and male PC1 scores (under the null hypothesis of no assortative mating) as follows:
Expected value = mean absolute distance between female PC1 score and PC1 scores all the possible combinations of males she might have mated with.
The **observed value** is the mean absolute distance between female PC1 score and PC1 score of all males mated with.

The above calculations are based on the total number of males a female actually mated with and the number of trials she took part in. The values are therefore different for each female because some did not take part in both trials (or did not spawn in both trials), and there was variation in the total number of males mated with over the course of the experiment (between 2-4).

For each female, the number of potential mates is a product of the possible combinations in trial 1 (T1) and trial 2 (T2). The number of combinations (choosing r males out of n males) in each trial are n!/r!(n-r)!. For example, the number of combinations is 7 for a female that mated with a single male in T1 and 588 for a female that mated with 2 males in T1 and 2 males in T2.

5. Measuring rhodopsin absorption spectra:

Reconstruction and measurement of absorption spectra of visual pigments:
Production, reconstruction, purification, and measurement of the visual pigments were performed as described Ueyama *et al*. (*66*) with minor modifications. Briefly, the sequences of rho (also known as RH1) H4 and H5 alleles were amplified by PCR using genomic DNA of Lake Massoko cichlids as a template with a pair of specific PCR primers (*27*) designed to produce a fusion protein with a FLAG-tag (Sigma-Aldrich) at its C terminus. The amplified DNA fragments were digested with restriction enzymes and cloned into the expression vector pFLAG-CMV-5a (Sigma-Aldrich). The visual pigments were reconstituted with A1-derived retinal. Absorption spectra of the pigment solutions in the presence of hydroxyl-amine (<100mM) before and after photobleaching were recorded using a spectrophotometer (UV-2400, Shimadzu, Japan). The measurements were taken 5 times before and after photobleaching. We determined the mean peak spectral values (maximum absorption spectra: λmax) and standard errors from multiple preparations and measurements for each pigment. All procedures after reconstitution of the pigments were performed under dim red light (>680 nm) conditions.

References

36. F. J. Rohlf, tpsDig2, (available at http://http//life.bio.sunysb.edu/morph/).

37. C. P. Klingenberg, MorphoJ: an integrated software package for geometric morphometrics. *Molecular ecology resources* (2011).

38. H. Li, Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv*. **q-bio.GN** (2013).

39. J. M. Catchen, A. Amores, P. Hohenlohe, W. Cresko, J. H. Postlethwait, Stacks: building and genotyping Loci de novo from short-read sequences. *G3 (Bethesda)*. **1**, 171–182 (2011).

40. T. M. Keane, C. J. Creevey, M. M. Pentony, Assessment of methods for amino acid matrix selection and their use on empirical data shows that ad hoc assumptions for choice of matrix are not justified. *BMC evolutionary …* (2006).

41. A. Stamatakis, RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics*. **22**, 2688–2690 (2006).

42. A. Stamatakis, P. Hoover, J. Rougemont, A rapid bootstrap algorithm for the RAxML Web servers. *Syst. Biol.* **57**, 758–771 (2008).

43. H. Shimodaira, M. Hasegawa, Multiple comparisons of log-likelihoods with applications to phylogenetic inference. *Mol. Biol. Evol.* (1999).

44. M. A. M. DePristo *et al.*, A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43**, 491–498 (2011).

45. A. McKenna *et al.*, The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).

46. H. Li, A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*. **27**, 2987–2993 (2011).

47. S. R. Browning, B. L. Browning, Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am. J. Hum. Genet.* **81**, 1084–1097 (2007).

48. S. Purcell *et al.*, PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *The American Journal of Human Genetics*. **81**, 559–575 (2007).

49. N. Patterson, A. L. Price, D. Reich, Population structure and eigenanalysis. *PLoS*

Genet. **2**, e190–e190 (2006).

50. O. Delaneau, J. Marchini, J.-F. Zagury, A linear complexity phasing method for thousands of genomes. *Nat. Methods*. **9**, 179–181 (2012).

51. O. Delaneau, B. Howie, A. J. Cox, J.-F. Zagury, J. Marchini, Haplotype estimation using sequencing reads. *Am. J. Hum. Genet.* **93**, 687–696 (2013).

52. R. R. R. Hudson, Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics*. **18**, 337–338 (2002).

53. G. Bhatia, N. Patterson, S. Sankararaman, A. L. Price, Estimating and interpreting F_{ST}: The impact of rare variants. *Genome Res.* **23**, 1514–1521 (2013).

54. A. R. Quinlan, I. M. Hall, BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*. **26**, 841–842 (2010).

55. J. Wakeley, Distinguishing migration from isolation using the variance of pairwise differences. *Theor Popul Biol*. **49**, 369–386 (1996).

56. M. Ashburner *et al.*, Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **25**, 25–29 (2000).

57. A. Alexa, J. Rahnenfuhrer, topGO: enrichment analysis for gene ontology. *R package version* (2010).

58. W. Huber *et al.*, Orchestrating high-throughput genomic analysis with Bioconductor. *Nat. Methods*. **12**, 115–121 (2015).

59. D. Merico, R. Isserlin, O. Stueker, A. Emili, G. D. Bader, Enrichment map: a network-based method for gene-set enrichment visualization and interpretation. *PLoS ONE*. **5**, e13984 (2010).

60. J. H. Morris *et al.*, clusterMaker: a multi-algorithm clustering plugin for Cytoscape. *BMC Bioinformatics*. **12**, 436 (2011).

61. L. Oesper, D. Merico, R. Isserlin, G. D. Bader, WordCloud: a Cytoscape plugin to create a visual semantic summary of networks. *Source Code Biol Med*. **6**, 7 (2011).

62. G. F. Turner, O. Seehausen, M. E. Knight, C. J. Allender, R. L. Robinson, How many species of cichlid fishes are there in African lakes? *Mol Ecol*. **10**, 793–806 (2001).

63. M. I. Taylor *et al.*, Characterization of tetranucleotide microsatellite loci in a Lake Victorian, haplochromine cichlid fish: a Pundamilia pundamilia x Pundamilia nyererei hybrid. *Mol Ecol Notes*. **2**, 443–445 (2002).

64. S. Gabriel, L. Ziaugra, D. Tabbaa, SNP genotyping using the Sequenom

MassARRAY iPLEX platform. *Curr Protoc Hum Genet*. **Chapter 2**, Unit 2.12 (2009).

65. X. Zheng *et al.*, A high-performance computing toolset for relatedness and principal component analysis of SNP data (2012).

66. H. Ueyama *et al.*, Novel missense mutations in red/green opsin genes in congenital color-vision deficiencies. *Biochem. Biophys. Res. Commun.* **294**, 205–209 (2002).

67. L. Ségurel, M. J. Wyman, M. Przeworski, Determinants of mutation rate variation in the human germline. *Annu Rev Genomics Hum Genet*. **15**, 47–70 (2014).

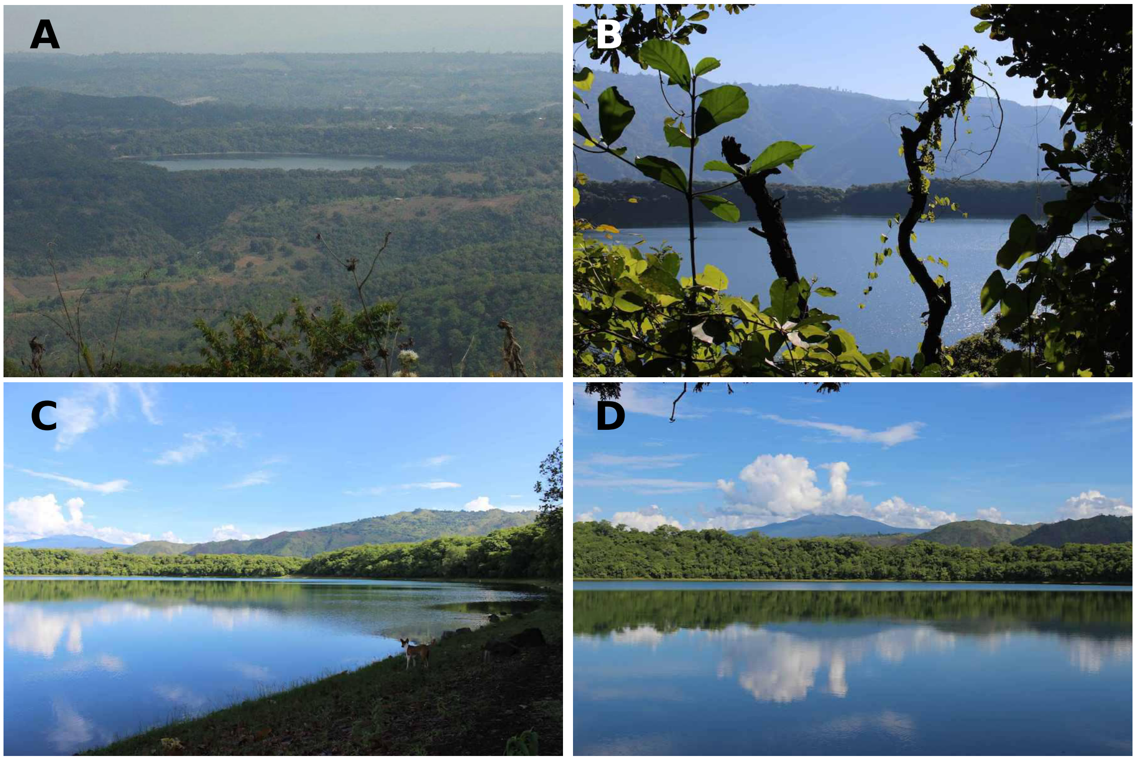


Fig. S1

Lake Massoko. Photographs of Lake Massoko taken from the: **(A)** east (photo taken from near to Lake Itende); **(B)** western crater rim looking east; **(C)** southern shore looking north-east; and **(D)** southern shore looking north.

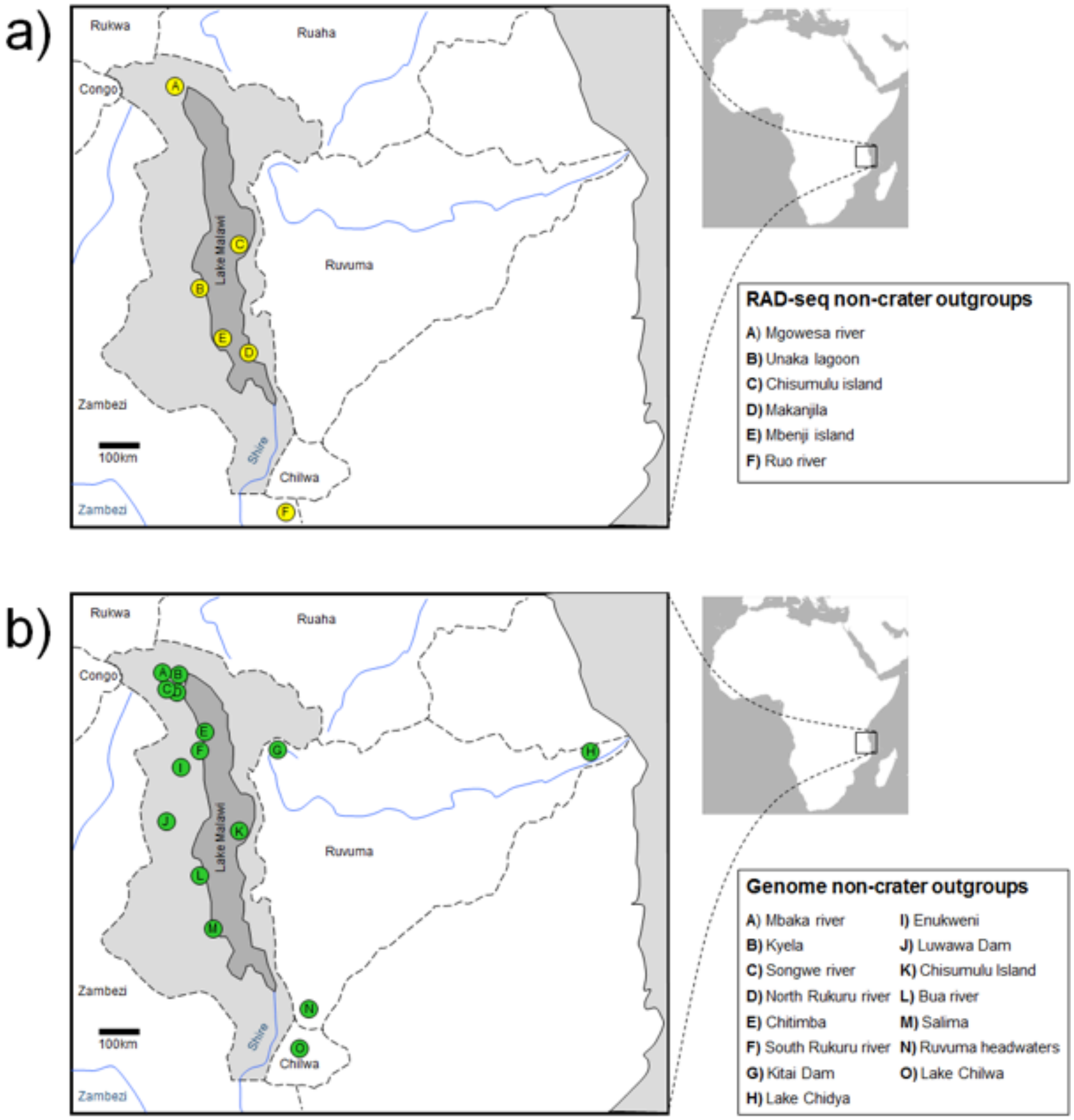


Fig. S2

Collection sites of non-crater-lake *Astatotilapia* calliptera specimens. Dotted lines represent catchment boundaries, with the Lake Malawi catchment shaded in gray. **a)** Specimens sequenced by RAD sequencing and used in phylogenetic analysis. **b)** Specimens for whole genome sequencing and comparisons with crater lake *Astatotilapia*.

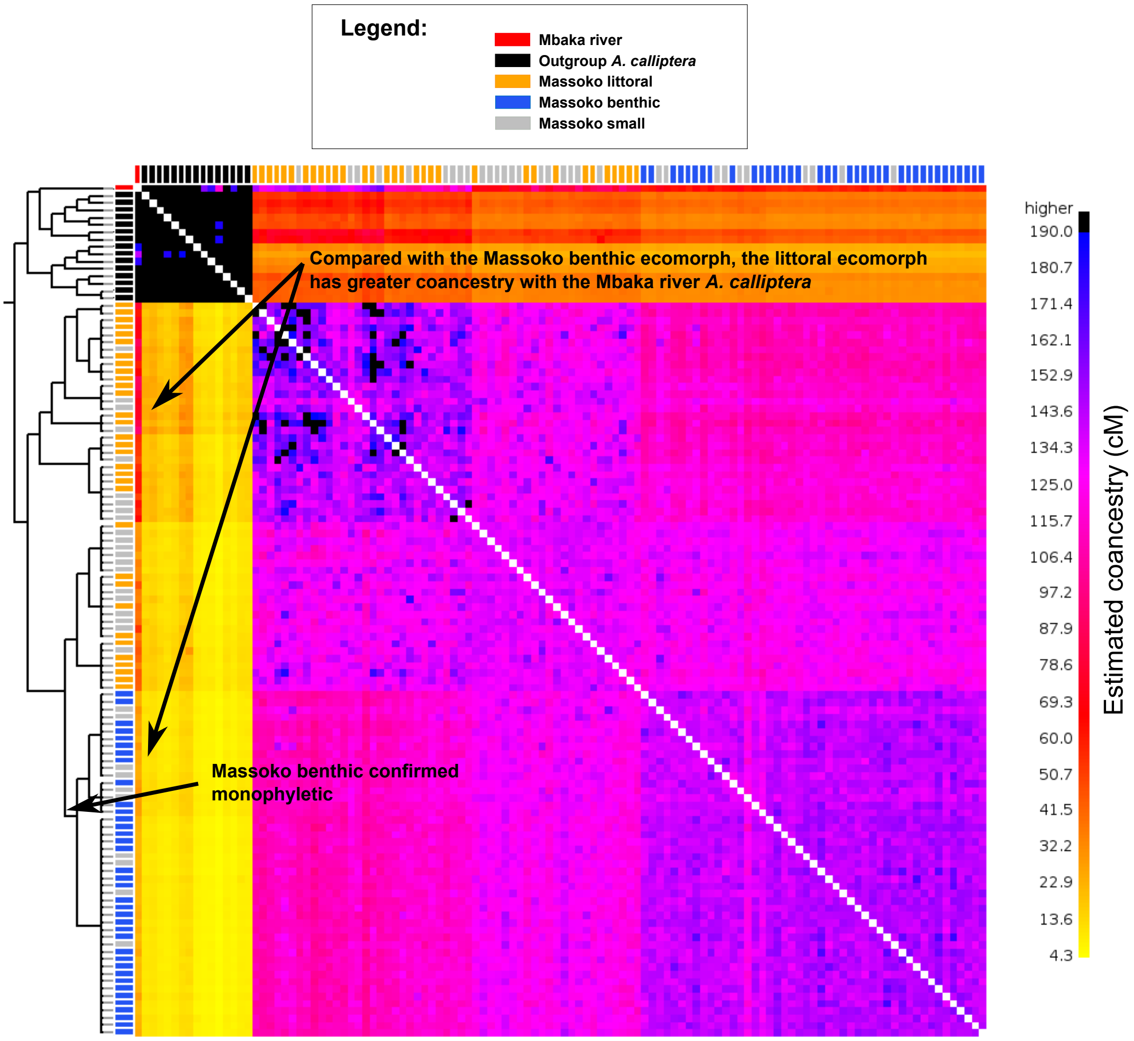


Fig. S3

fineSTRUCTURE results. Co-ancestry matrix with the tree showing inferred relationships between samples. Each tip and label correspond to an individual, with labels colored according to the population/ecomorph as indicated in the legend. The results show tight clustering and monophyly of the benthic ecomorph, greater population structure within the littoral ecomorph, and a difference between the ecomorphs with respect to co-ancestry with Mbaka river *A. calliptera*, as indicated.

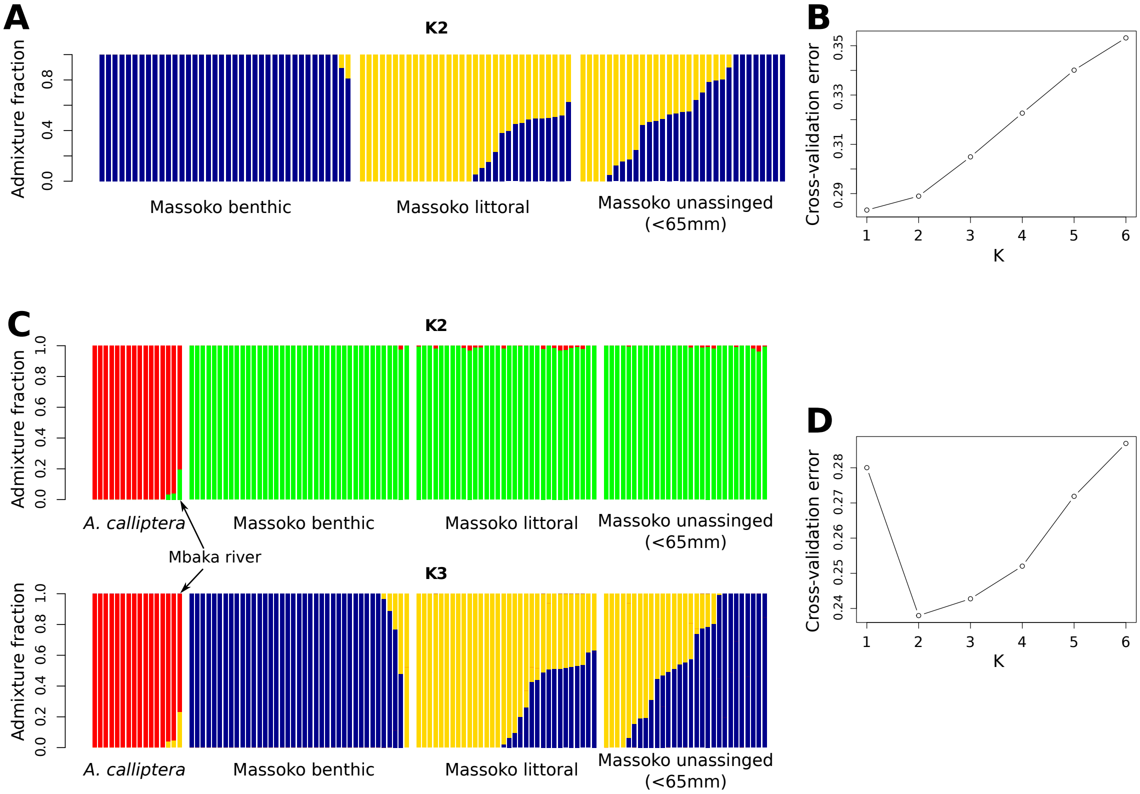


Fig. S4

ADMIXTURE estimates of individual ancestries. (A-B) Massoko samples only: **(A)** With two postulated ancestral populations (`K=2`), benthic individuals form a virtually homogenous group. Eleven of the samples field-assigned as littoral appear to be >25% admixed. The unassigned samples are a mixture of benthic, littoral, and admixed individuals. **(B)** ADMIXTURE cross-validation approach to choosing the `K` parameter - the postulated number of ancestral populations. The error estimates are based on 10-fold cross-validation. The lowest error is observed with `K=1`, suggesting that population differentiation between the ecomorphs is subtle (*6*, *13*). **(C-D)** Including additional *A. calliptera* samples: **(C)** A little gene-flow from A. calliptera into Massoko is apparent with `K=2`, but not with `K=3`. There appears to be gene flow out of Massoko into Mbaka river, possibly stronger than the inward gene flow into Massoko. **(D)** Cross-validation results.

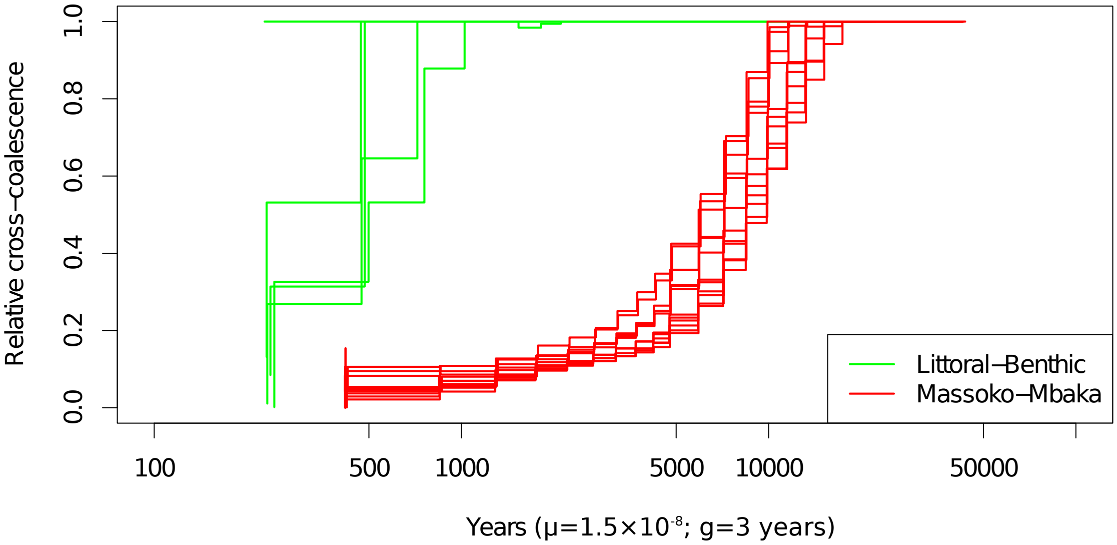


Fig. S5

MSMC cross-coalescence between littoral and benthic ecomorphs (green) and between all Massoko high-coverage individuals and the sample from Mbaka river. MSMC infers effective coalescence rates as a function of time: across the two populations and within populations. The 'relative cross-coalescence rate' (y axis) is a measure for the genetic separation of populations, constituting the ratio between the cross-population and within-population coalescence rates: the rate should be close to 1 when the two populations are well mixed and 0 after they have fully separated (*17*). The split of Massoko individuals from the Mbaka river sample (red) is approximately ten times earlier than any separation between benthic and littoral ecomorphs is observed (green). The time axis values assume: 1) average generation time g=three years and 2) per generation mutation rate μ=1.5x10^{-8}, making the assumption that the μ in cichlids is similar to μ estimated in human studies (*67*). Direct estimate of μ in cichlids is not available.

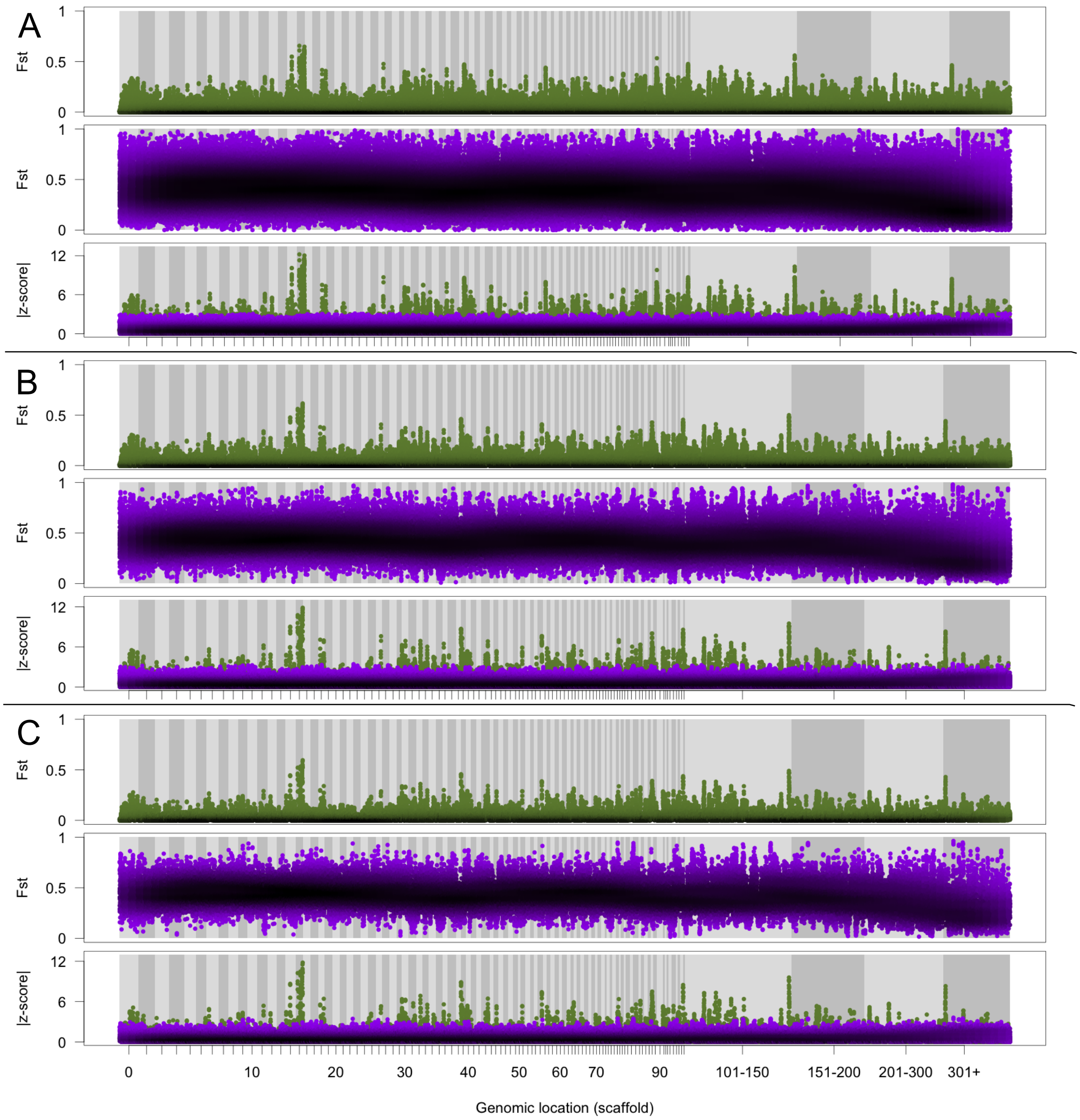


Fig. S6

Genome-wide pattern of F_{ST} divergence using sliding windows or varying sizes. The overall pattern of ‘genomic islands’ raising above low background divergence is unaffected by varying the window size. Each figure shows genome-wide pattern of F_{ST} between Massoko benthic and Massoko littoral (green), and between combined Massoko and Itamba populations (purple), and absolute z-scores of Massoko- Itamba divergence (purple) and within-Massoko divergence (green). **(A)** Window size=15 variants; **(B)** Window size=50 variants; **(C)** Window size=100 variants.

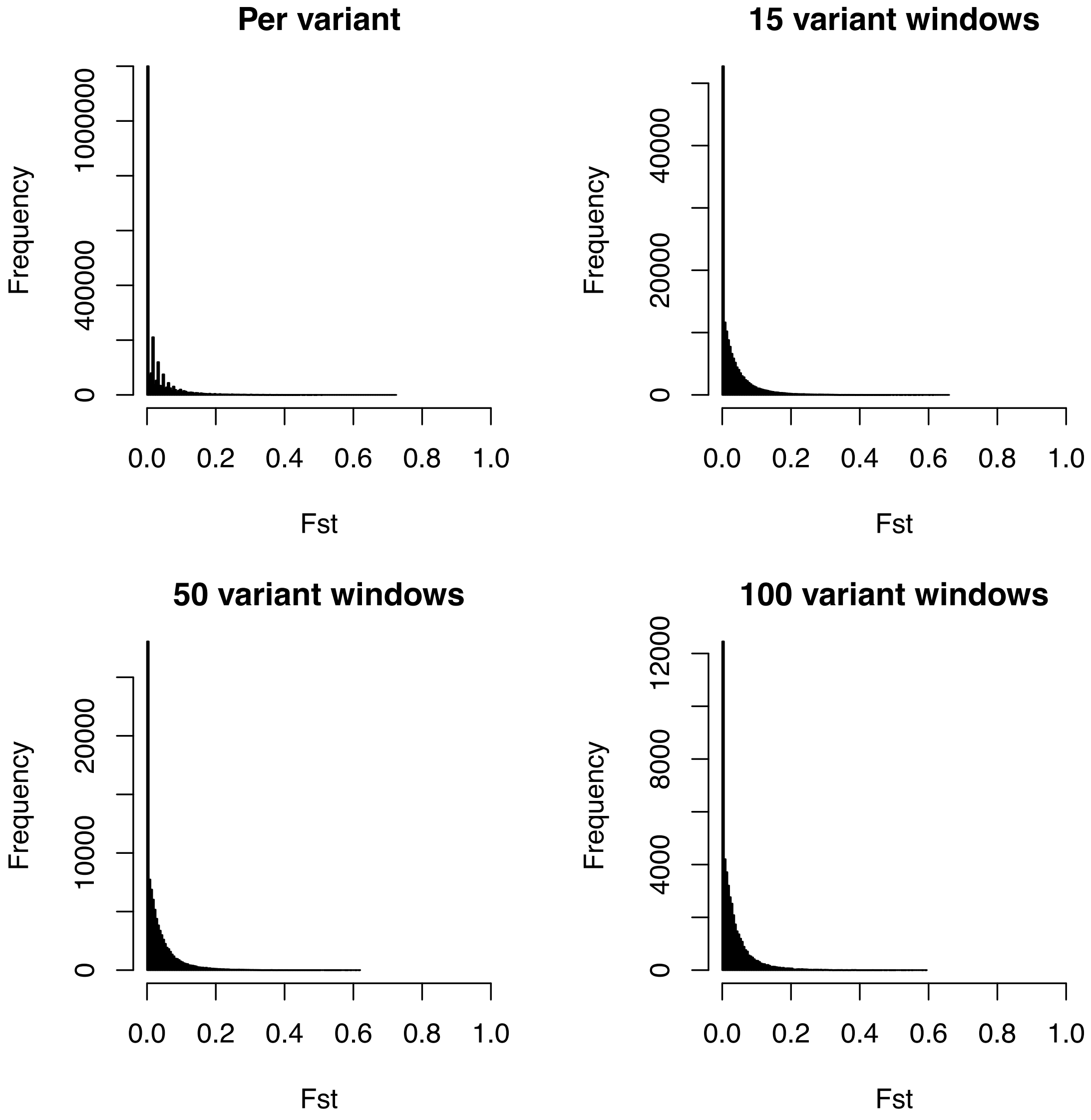


Fig. S7

Statistical distribution of within-Massoko F_{ST} divergence, per variant and in sliding windows of varying sizes. The distribution has a sharp L like shape, largely independent of the window size used, consistent with theoretical predictions about early stages of speciation with gene-flow (*2*).

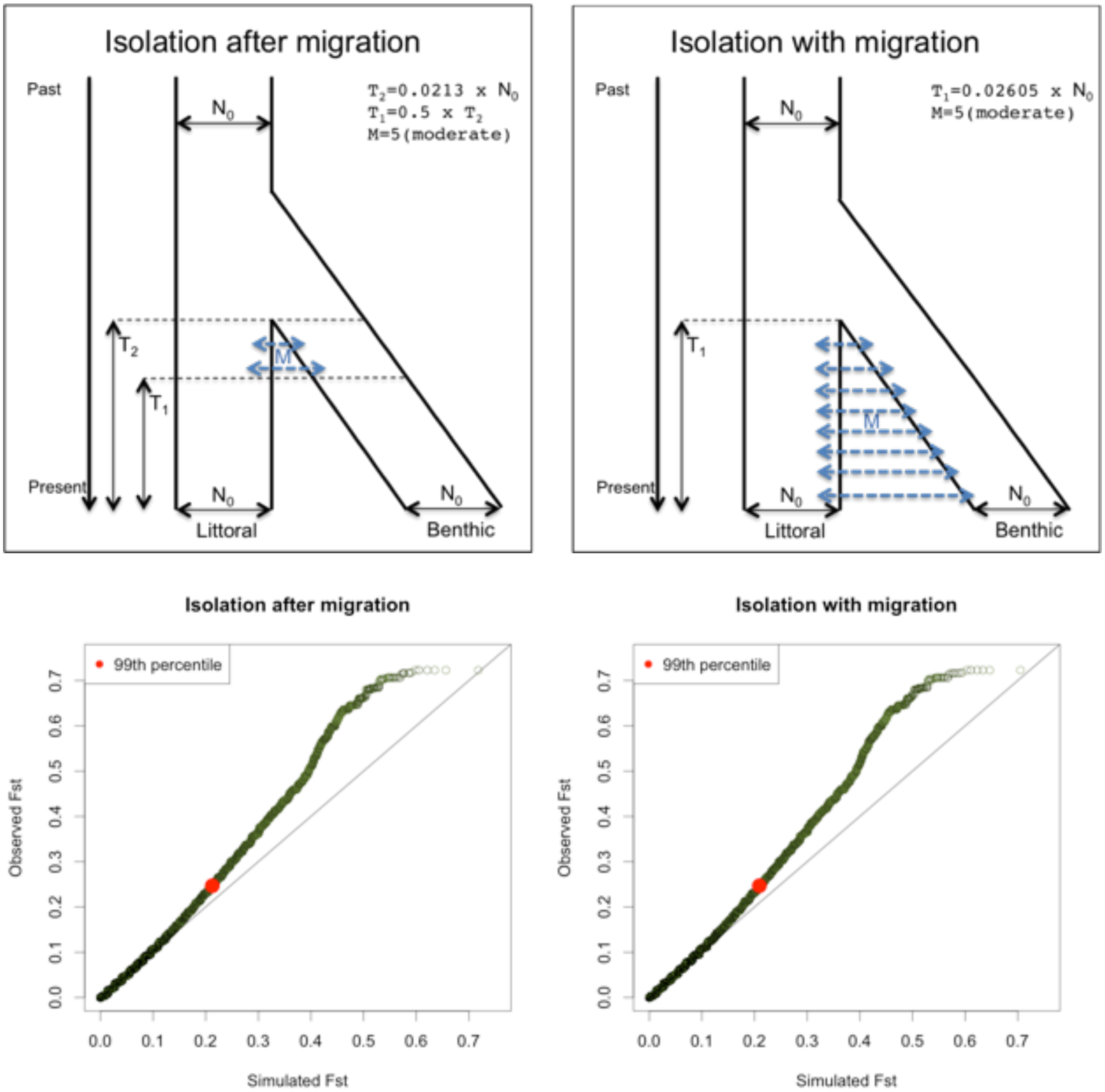


Fig. S8

Two basic models of species formation used for coalescent simulations (top) and quantile-quantile plots comparing the distributions of simulated F_{ST} values with observed benthic-littoral divergence (bottom). Darker color indicates greater density of datapoints, and the position of the 99th percentile of both distributions is indicated in red. Similar patterns are observed for simulations under both models.

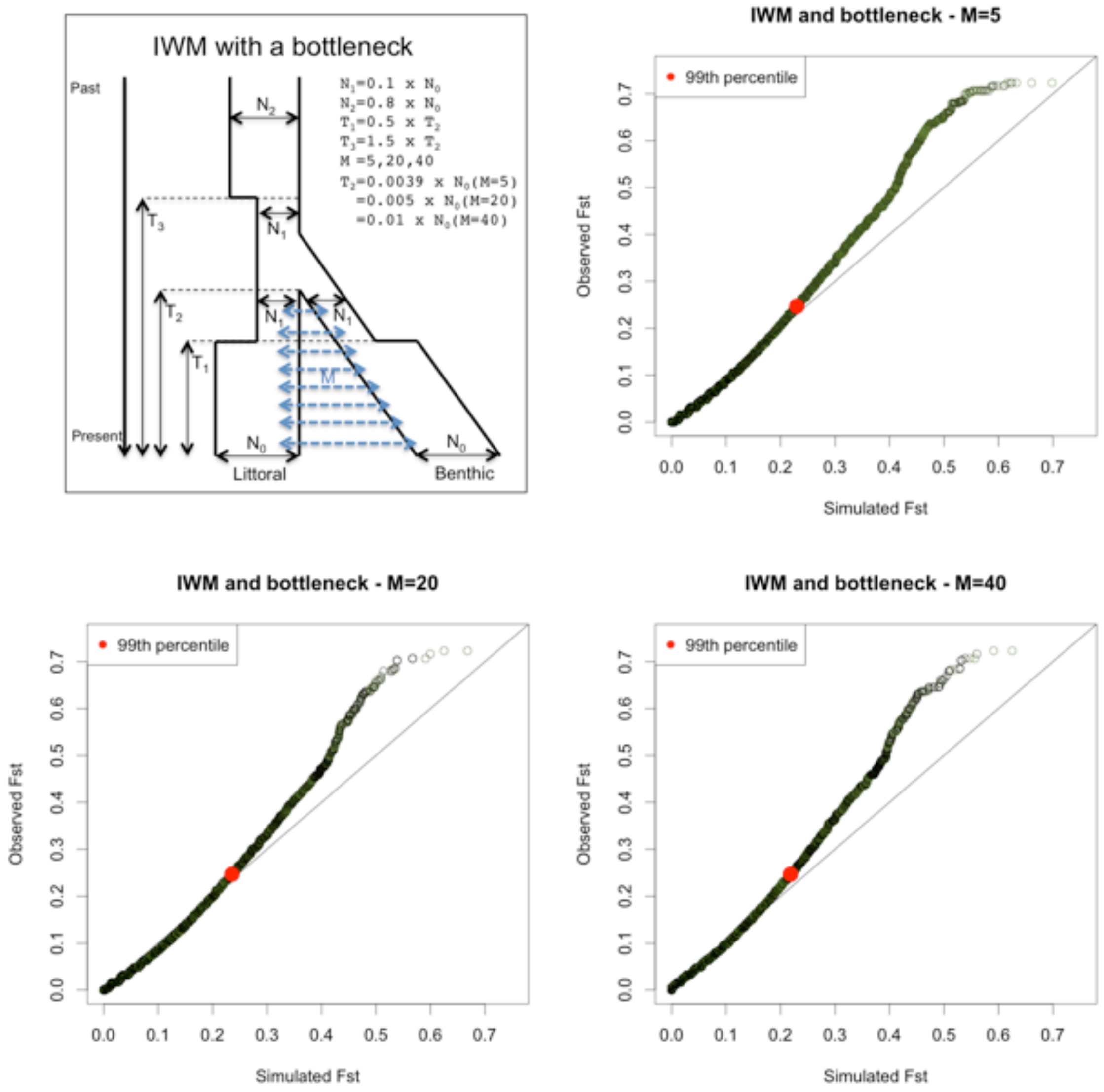


Fig. S9

The isolation with migration (IWM) model of species formation with a strong population bottleneck (top left) was used for neutral coalescent simulations, with a range of values for the migration parameter M. Quantile-quantile plots compare the distributions of simulated F_{ST} values under this model with observed benthic-littoral divergence (top-right and bottom). Darker color indicates greater density of datapoints, and the position of the 99th percentile of both distributions is indicated in red. Similar patterns are observed for all simulations indicating approximately the top 1% of observed values are higher than corresponding simulated values.

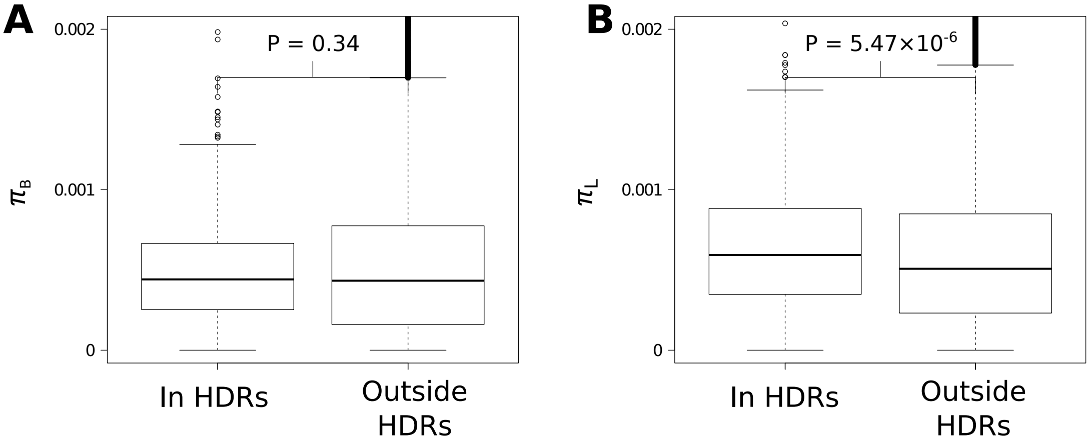


Fig. S10

Comparison of nucleotide diversity in HDRs with the rest of the genome: (A) in the benthic morph (π_B); **(B)** in the littoral morph (π_L).

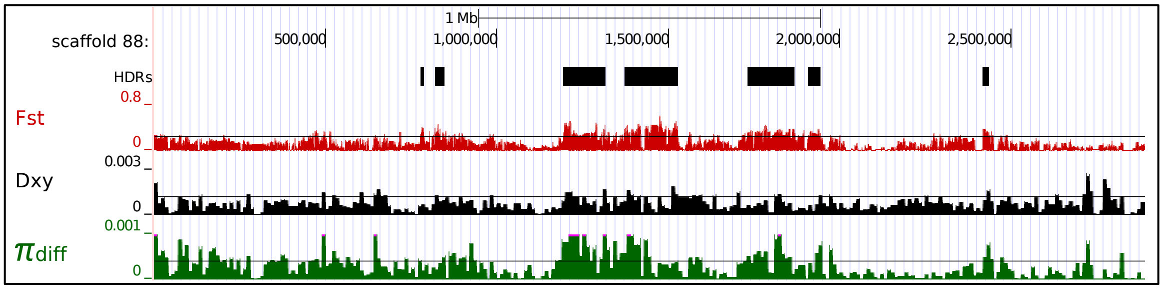


Fig. S11

Patterns of F_{ST} , d_{XY} , and π_{diff} in a speciation cluster on scaffold 88. In this one region of the genome, F_{ST} appears to be elevated above or close to the 99th percentile (black line, representing maximum observed neutral divergence in simulations) over the distance of 2Mb, suggesting that divergence hitchhiking might have started forming a broader 'continent' of divergence.

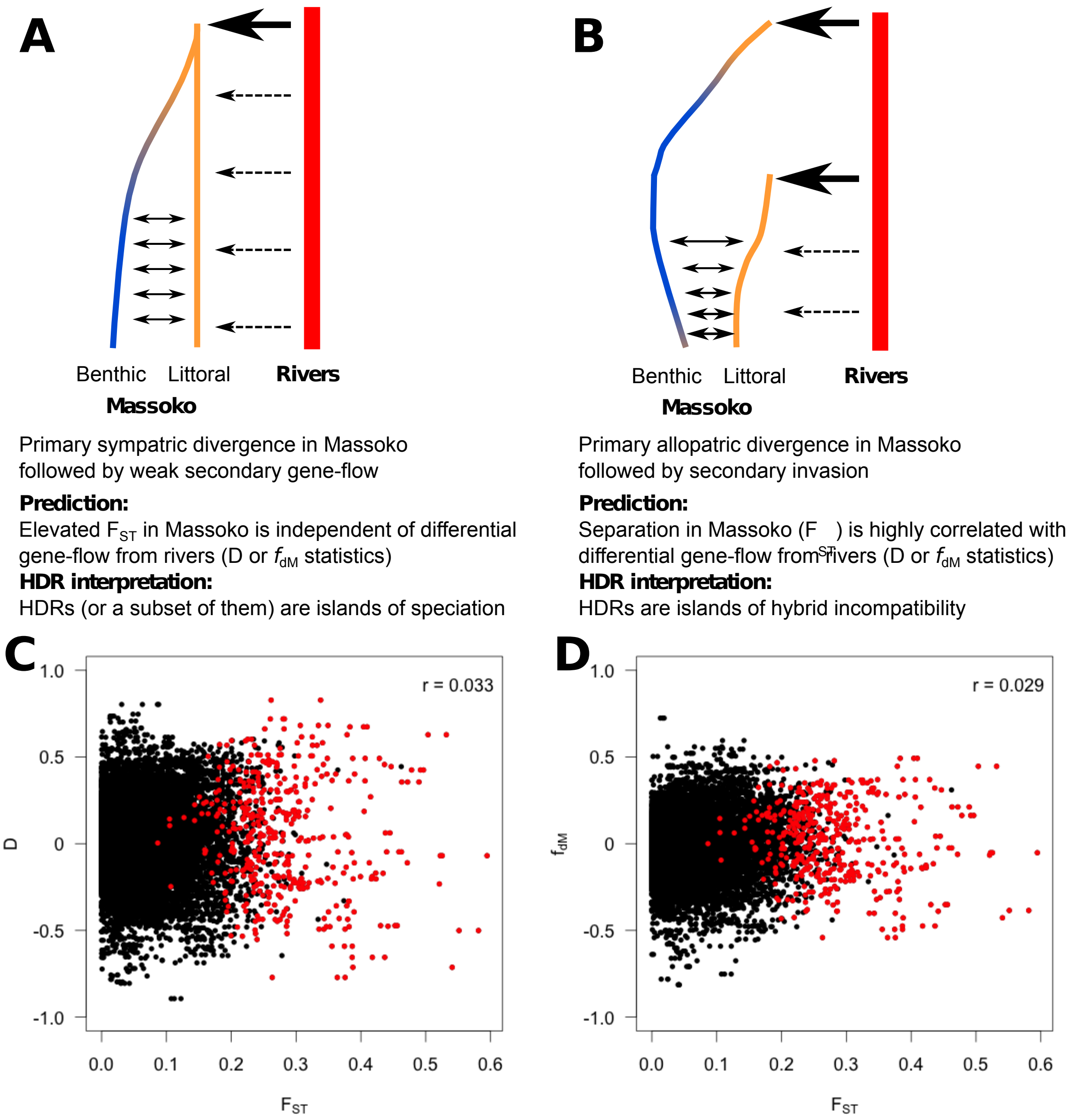


Fig. S12

Evidence against allopatric (double-invasion) divergence between benthic and littoral ecomorphs. (A-B) Two geographic models of divergence with predictions and interpretation of HDRs. **(A)** Primary sympatric divergence in Massoko. **(B)** Primary allopatric divergence. **(C-D)** F_{ST} between benthic and littoral ecomorphs is independent of levels of differential gene-flow measured by **(C)** Patterson's D, and **(D)** the f_{dM} statistic for locating introgressed regions. F_{ST} was averaged over windows with 15 variants, and D and f_{dM} were averaged over windows of 50 informative variants (i.e. variants with ABBA or BABA patterns). Values for windows within HDRs are shown in red, values for windows outside HDRs are shown in black. Pearson correlation coefficients r are displayed in the top right corner of each figure. The lack of correlation between either D or f_{dM} with F_{ST} is consistent with the predictions of the model of primary sympatric divergence.

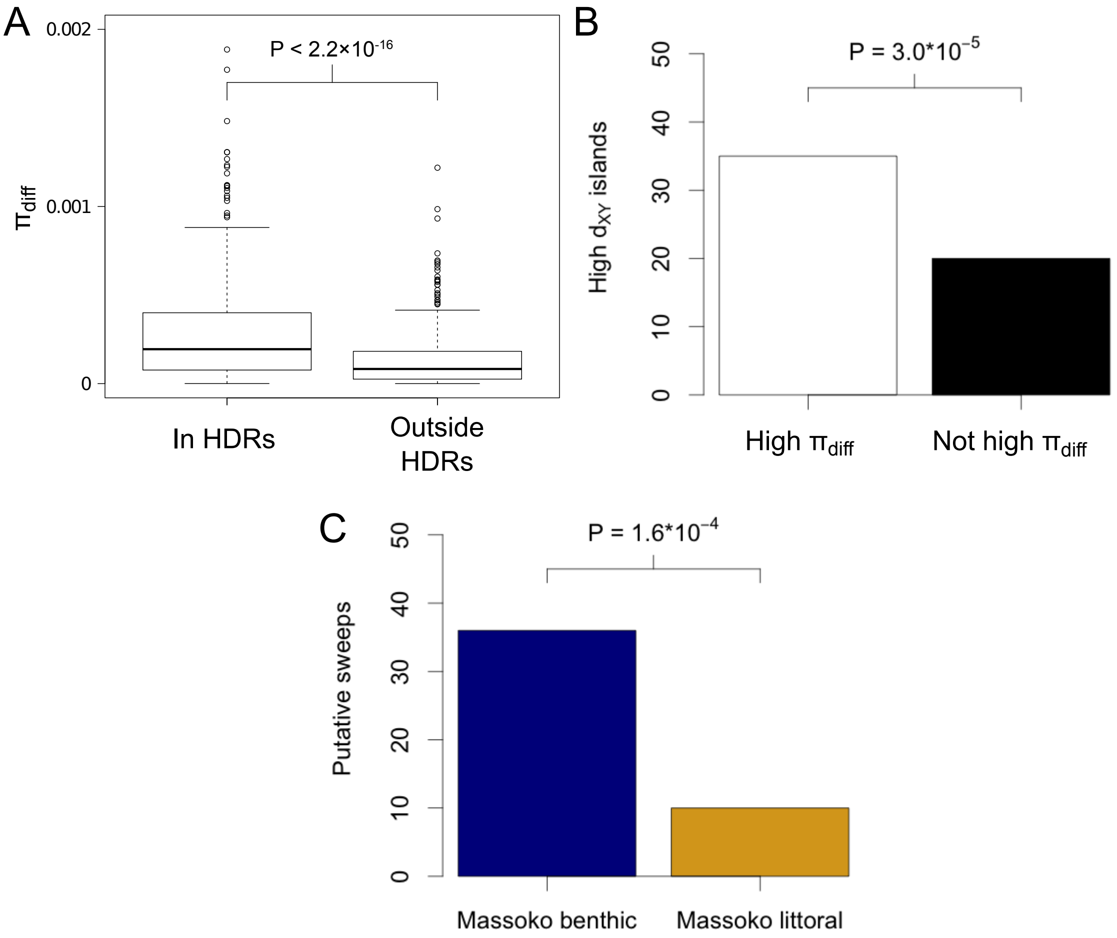
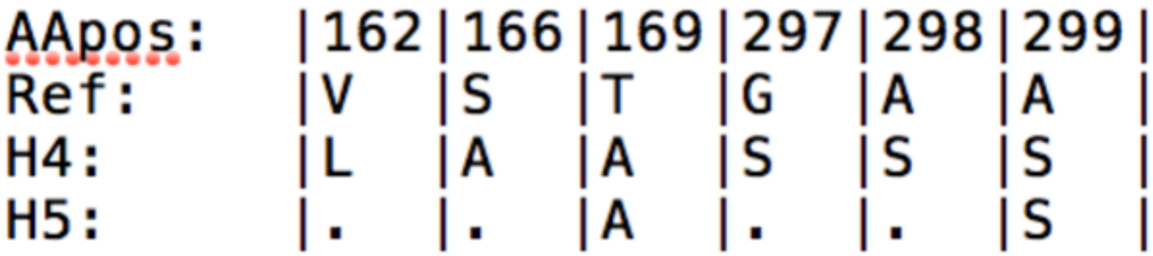


Fig. S13

Characterization of HDRs in terms of the magnitude of difference in nucleotide diversity between benthic and littoral ecomorphs (π_{diff}). (A) π_{diff} is significantly higher within HDRs ($P < 2.2\times10^{-16}$, two-tailed Mann-Whitney test), compared with the rest of the genome. **(B)** The overlap between HDRs with high d_{XY} (putative 'speciation islands') and HDRs with high π_{diff} (putative recent selective sweeps) is significant. Thirty-five out of 55 high d_{XY} islands also have high π_{diff} ($P=3\times10^{-5}$, hypergeometric test). **(C)** Thirty-six out of the 46 putative selective sweeps are in the Massoko benthic morph, providing significant evidence that positive selection has been more prevalent in the benthic form ($P<1.6\times10^{-4}$, two tailed Binomial test).



```
AApos:   |162|166|169|297|298|299|
Ref:     |V  |S  |T  |G  |A  |A  |
H4:      |L  |A  |A  |S  |S  |S  |
H5:      |.  |.  |A  |.  |.  |S  |
```

Fig. S14

Amino-acid differences between the two haplotypes of the rhodopsin (*rho*) found in Lake Massoko *Astatotilapia* are present at positions 162, 166, 297, and 298. There are two additional amino-acid positions (169, 299) where both ecomorphs differ from the *M. zebra* reference.

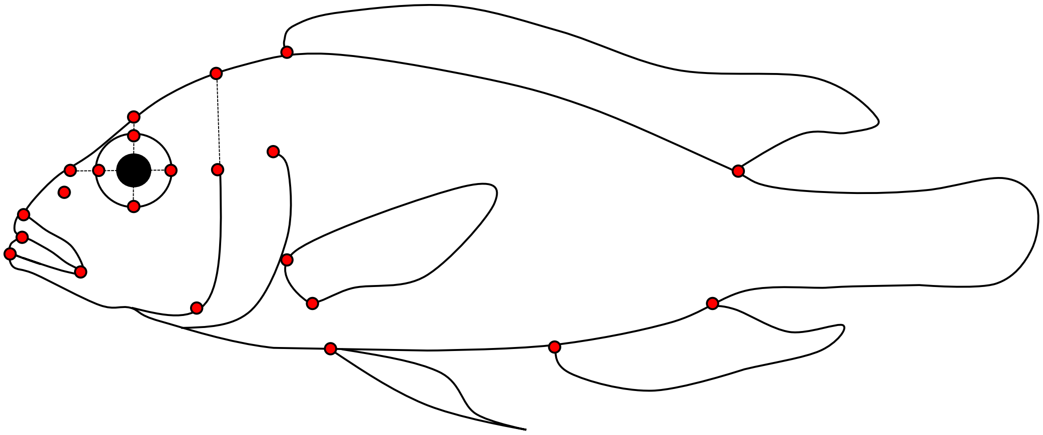


Fig. S15
The 22 landmarks used for quantifying variation in head and body morphology.

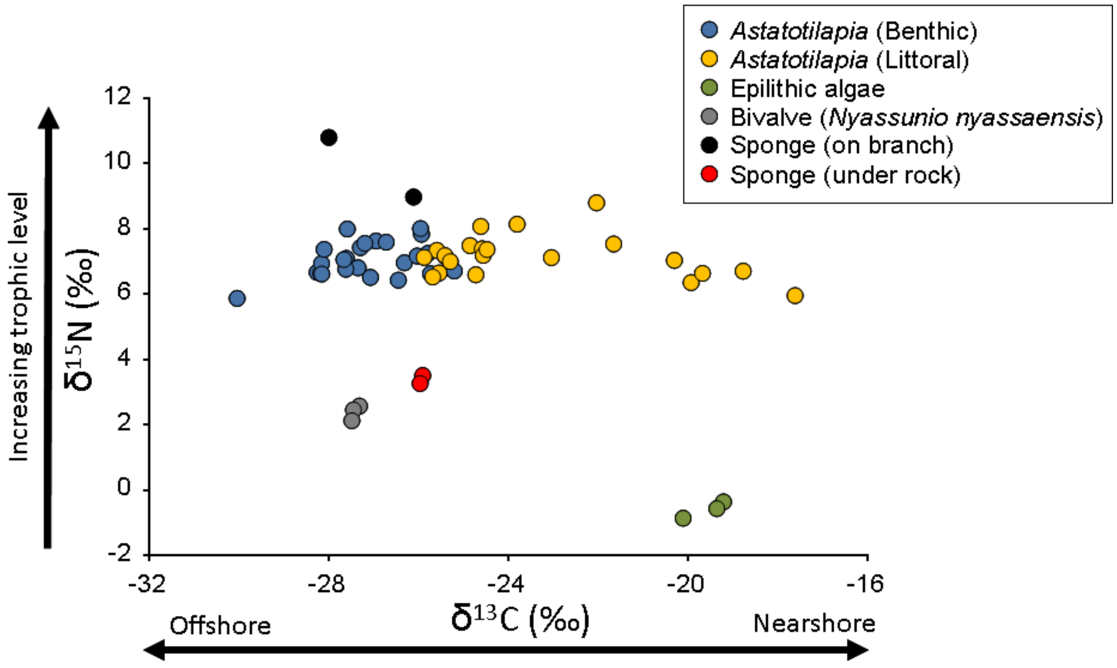


Fig. S16

Results of stable isotope analysis of the Lake Massoko *Astatotilapia* ecomorphs, together with environmental samples from Massoko.

Table S1

Results of a survey of fish fauna in six crater lakes of Rungwe District, Tanzania, conducted in July and November 2011.

Lake	Species	Family	Tribe	Probable Status
Kingiri	*Astatotilapia sp.* 'kingiri black'	Cichlidae	Haplochromini	Endemic
	Rhamphochromis sp. 'kingiri dwarf'	Cichlidae	Haplochromini	Endemic
	Rhamphochromis sp. 'kingiri brevis'	Cichlidae	Haplochromini	Endemic
	Serranochromis robustus	Cichlidae	Haplochromini	Native
	Coptodon rendalli	Cichlidae	Tilapiini	Native
	Oreochromis shiranus	Cichlidae	Tilapiini	Native
	Oreochromis (Nyasalapia)	Cichlidae	Tilapiini	Native
	squamipinnis			Native
	Clarias gariepinus	Clariidae		Native
	Micropanchax johnstoni	Poeciliidae		Native
	Barbus radiatus	Cyprinidae		Native
	Barbus trimaculatus	Cyprinidae		
Ilamba	*Astatotilapia sp.* 'ilamba black'	Cichlidae	Haplochromini	Endemic
	Otophraynx sp. 'Ilamba tetrastigma'	Cichlidae	Haplochromini	Endemic
	Oreochromis cf shiranus	Cichlidae	Tilapiini	Native/Endemic?
	Oreochromis cf squamipinnis	Cichlidae	Tilapiini	Native/Endemic?
	Clarias gariepinus	Clariidae		Native
	Mesobola cf. spinifer	Cyprinidae		Native
	Barbus paludinosos	Cyprinidae		Native
	Barbus trimaculatus	Cyprinidae		Native
	Barbus macrotaenia	Cyprinidae		Native
	Barbus radiatus	Cyprinidae		Native
Ikapu	*Astastotilapia sp.* 'ikapu dark'	Cichlidae	Haplochromini	Endemic
	Tilapia sparrmanii	Cichlidae	Tilapiini	Native/Introduced?
	Oreochromis 'golden chambo'	Cichlidae	Tilapiini	Endemic
	Clarias gariepinus	Clariidae		Native/Introduced?
Itamba	*Astatotilapia sp.* 'itamba dark'	Cichlidae	Haplochromini	Endemic
	Oreochromis cf. shiranus	Cichlidae	Tilapiini	Native/Endemic?
	Oreochromis (Nyasalapia) cf. karongae	Cichlidae	Tilapiini	Native/Endemic?
	Oreochromis niloticus	Cichlidae	Tilapiini	Introduced (may not be established)
Massoko	*Astatotilapia sp.* 'massoko benthic'	Cichlidae	Haplochromini	Endemic
	Astatotilapia sp. 'massoko littoral'	Cichlidae	Haplochromini	Endemic
	Coptodon rendalli	Cichlidae	Tilapiini	Introduced?
	Oreochromis (Nyasalapia) squamipinnis	Cichlidae	Tilapiini	Native/Endemic?
	Clarias gariepinus	Clariidae		Introduced?
Itende	*Astatotilapia sp.* 'itende'	Cichlidae	Haplochromini	Endemic
	Oreochromis (Nyasalapia) spp. squamipinnis	Cichlidae	Tilapiini	Endemic/Native?

Table S2

Location and geographical characteristics of crater lakes with haplochromine cichlid fauna in Rungwe District, Tanzania. Data from (*8*), except Ikapu, estimated from Google Earth and own survey of depth.

	Latitude	Longitude	Altitude	Surface area	Max. depth	Volume
Ikapu	9°22' S	33°48' E	653 m	0.28 km^2	3 m	$0.85x10^6$ m^3
Ilamba	9°24' S	33°50' E	548 m	0.42 km^2	23 m	$7.01x10^6$ m^3
Itamba	9°21' S	33°51' E	821 m	0.12 km^2	18 m	$0.69x10^6$ m^3
Itende	9°19' S	33°47' E	1020 m	0.14 km^2	2 m	$0.28x10^6$ m^3
Kingiri	9°25' S	33°51' E	515 m	0.27 km^2	34 m	$5.37x10^6$ m^3
Massoko	9°20' S	33°45' E	845 m	0.38 km^2	37 m	$8.91x10^6$ m^3

Table S3

Depth distribution of ecomorphs in Lake Massoko. Based on collections using a variety of methods (*6*) in July-August and December 2014, and August 2015. There is a significant association between bottom depth and morph frequencies ($\chi^2_{4\,df}$= 207.1, P<0.001).

	0-5m	5-10m	10-15m	15-20m	20-25m	Total
Benthic	0	6	11	25	75	117
Littoral	98	54	15	21	0	188
Total	98	60	26	46	75	305
% Benthic	0	10	42.3	54.3	100	
% Littoral	100	90	57.7	45.7	0	

Table S4
An overview of *Astatotilapia* samples collected for RAD sequencing.

Sampling location (ecomorph)	N	Sampling Dates	Collector(s)	Latitude S	Longitude E
Lake Massoko (benthic)	5	17/07/2011	MG, BN, GT, SM, AS	9°20'0	33°45'18
Lake Massoko (littoral)	3	17/07/2011	MG, BN, GT, SM, AS	9°20'0	33°45'18
Lake Massoko (small, unassigned)	4	17/07/2011	MG, BN, GT, SM, AS	9°20'0	33°45'18
Lake Itende	3	27/11/2011	MG, GT, AS	9°19'19	33°47'15
Lake Ikapu	3	20/07/2011	MG, BN, GT, SM, AS	9°22'12	33°48'25
Lake Itamba	2	19/07/2011	MG, BN, GT, SM, AS	9°21'04	33°50'39
Lake Ilamba	2	17/07/2011	MG, BN, GT, SM, AS	9°23'33	33°50'09
Lake Kingiri	8	15+21/07/2011	MG, BN, GT, SM, AS	9°25'08	33°51'29
Ruo river	2	22/5/2009	MG, AS, JS	15°50'77	35°11'69
Unaka lagoon	1	24 /07/2011	MG, AS	12°23'59	34°05'17
Mbenji island	2	--/01/2011	MG (from imported wild stock)	~13°26'	~34°29'
Makanjila	2	18 /01/ 2011	MG, PP, JB	13°41'35	34°50'51
Chisumulu island	1	--/01/2011	MG (from imported wild stock)	~12°00'	~34°37'
Mgowesa river	2	16/07/2011	MG, BN, GT, SM, AS	9°23'43	33°49'38
***Astatotilapia tweddlei* (Lake Chilwa)**	1	19/05/2009	MG, AS, JS 305A	15°22'18	35°35'20

MG = Martin Genner, GT = George Turner, BN = Benjamin Ngatunga, SM = Semvua Mzighani, AS = Alan Smith, JS = Jennifer Swanstrom, PP = Paul Parsons, JB = Jon Bridle.

Table S5

Results of morphological and stable isotope analysis. Analysis of Covariance (ANCOVA) tests of morphological and stable isotope differences among benthic and littoral morphs. In each case total length (TL) was employed as a covariate.

	N benthic	*N* littoral	*F* TL	*P* TL	*F* ecomorph	*P* ecomorph
External morphology (PC1)	41	73	5.749	0.018	166.884	< 0.001
Body mass*	15	19	677.780	< 0.001	34.170	< 0.001
Pharyngeal jaw mass*	15	19	110.432	< 0.001	18.337	< 0.001
Pharyngeal jaw tooth width	15	19	4.037	0.053	25.121	< 0.001
Stable isotopes (δ^{13}C)	24	22	3.296	0.076	46.834	< 0.001
Stable isotopes (δ^{15}N)	24	22	0.516	0.476	0.636	0.430

*$\log_{10}$ transformed

Table S6

An overview of *Astatotilapia* samples collected for whole genome sequencing.

Sampling location (ecomorph)	N	Sequencing: ~coverage/ chemistry	Sampling Dates	Collector(s)	Latitude S	Longitude E
Lake Massoko (benthic)	6 31	15x/v3 6x/v4	23-24/11/2011	MG, GT, BN, SM, AS	9°20’0	33°45’18
Lake Massoko (littoral)	6 26	15x/v3 6x/v4	17/7/2011; 21-23/11/2011	MG, GT, BN, SM, AS	9°20’0	33°45’18
Lake Massoko (small unassigned)	31	6x/v4	17/7/2011, 23-25/11/2011	MG, GT, BN, SM, AS	9°20’0	33°45’18
Lake Itamba	30	6x/v4	19/7/2011, 22/11/2011, Lab stock, collected wild Nov 2011.	MG, GT, BN, SM, AS	9°21’04	33°50’39
Chitimba	1	15x/v4	27/01/2014	HS	10°34’37	34°10’14
North Rukuru	1	15x/v4	28/01/2014	HS	9°55’01	33°55’39
Songwe River	1	15x/v4	28/01/2014	HS	9°35’14	33°46’10
South Rukuru	1	15x/v4	19/10/2013	HS	10°45’42	34°07’33
Enukweni	1	15x/v4	Lab stock, origins 2004	MG, (from import wild stock)	11°11’14	33°52’52
Lake Chidya	1	15x/v4	18/08/2013	MG, BN, SM, AS	10°35.49	40°9’19
Kitai Dam	1	15x/v4	06/09/2012	MG, GT, BN, SM, AS	10°42’22	35°11’46
Ruvuma river	1	15x/v3	17/5/2009	MG, JS	14.22’22	35.32’54
Near Kyela	1	15x/v4	14/07/2011	MG, GT, BN, SM, AS	9°33’05	33°53’11
Luwawa Dam	1	15x/v4	29/05/2010	JS	12°06’57	33°43’23
Bua	1	15x/v4	13/09/2012	MG, AS	13°18’30	33°32’51
Chisumulu island	1	15x/v3	23/09/2012	PP, JS	~12°00’	~34°37’
Mbaka River	1	15x/v3	17/7/2011	MG, GT, BN, SM, AS	9°20’27	33°47’04
Salima	2	15x/v3	Lab stock	AT, GT	~13°46’	~34°27’
Lake Chilwa	1	15x/v3	19/9/2012	PP, JS	15°22’15	35 °35’30

MG = Martin Genner, GT = George Turner, BN = Benjamin Ngatunga, SM = Semvua Mzighani, AS = Alan Smith, Jennifer Swanstrom, PP Paul Parsons, HS = Harold Sungani

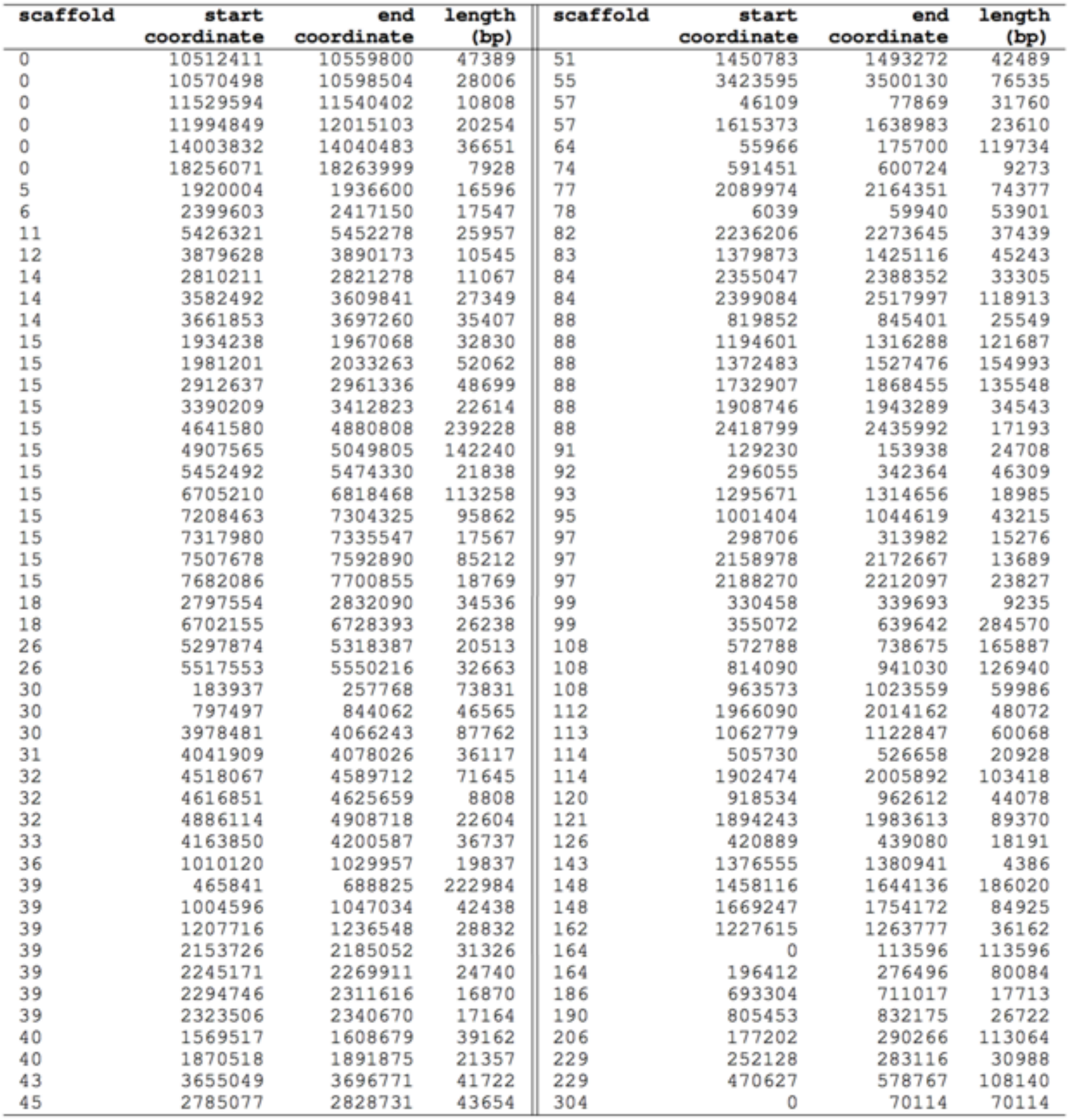
Table S7

A summary of sliding-window based F_{ST} calculations for Massoko benthic-littoral divergence.

Window size (variants)	Average length (bp)	F_{ST} range	Median F_{ST}	Proportion with zero F_{ST}	95th percentile	99th percentile
1	NA	0.00 - 0.72	0.003	0.476	0.126	0.247
15	5,369	0.00 - 0.66	0.016	0.258	0.134	0.240
50	17,839	0.00 - 0.62	0.018	0.208	0.129	0.231
100	35,455	0.00 - 0.60	0.019	0.171	0.126	0.225
500	174,390	0.00 - 0.46	0.024	0.064	0.115	0.197

Table S8

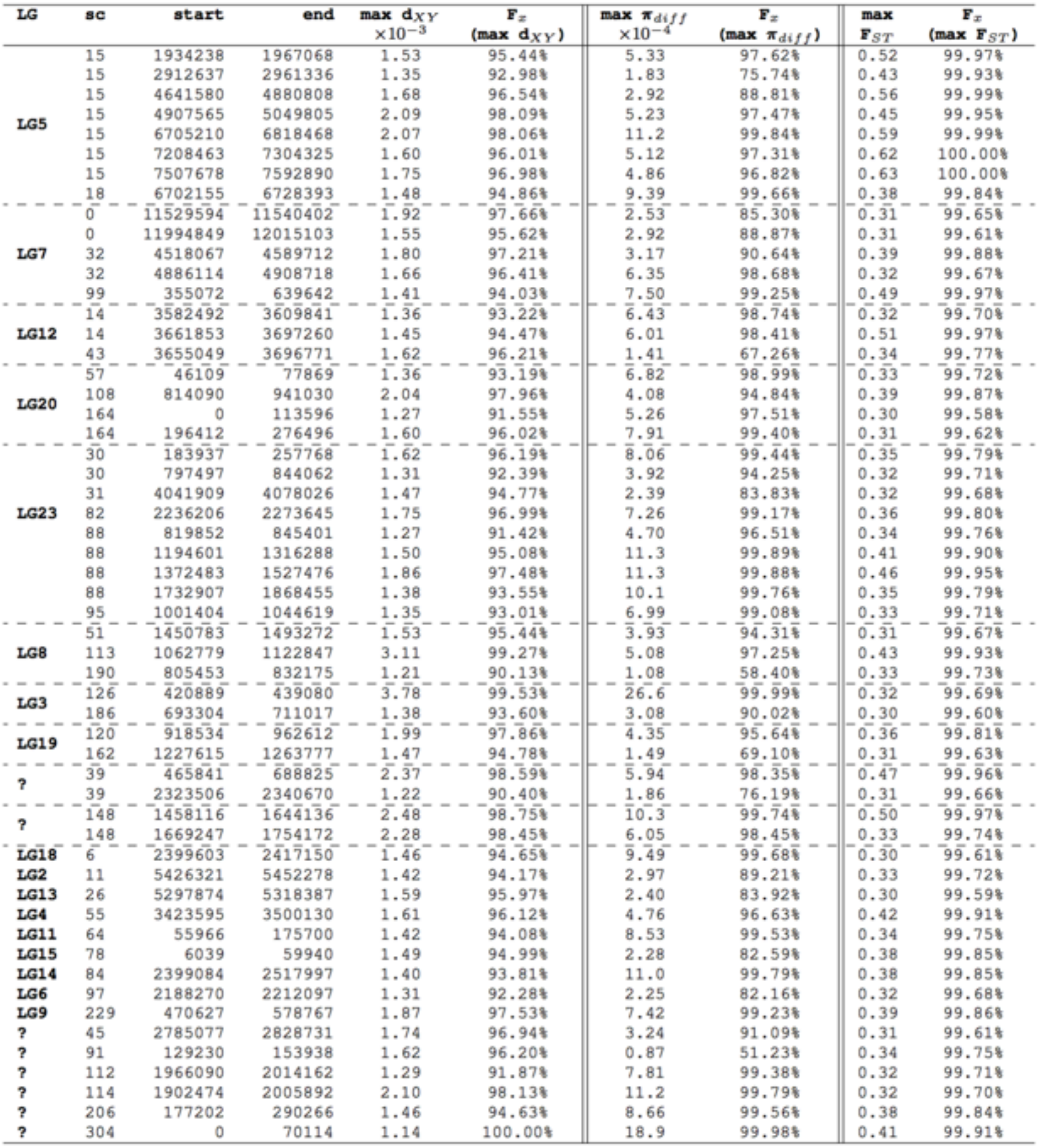
Genomic location and lengths of highly diverged regions (HDRs).



scaffold	start coordinate	end coordinate	length (bp)
0	10512411	10559800	47389
0	10570498	10598504	28006
0	11529594	11540402	10808
0	11994849	12015103	20254
0	14003832	14040483	36651
0	18256071	18263999	7928
5	1920004	1936600	16596
6	2399603	2417150	17547
11	5426321	5452278	25957
12	3879628	3890173	10545
14	2810211	2821278	11067
14	3582492	3609841	27349
14	3661853	3697260	35407
15	1934238	1967068	32830
15	1981201	2033263	52062
15	2912637	2961336	48699
15	3390209	3412823	22614
15	4641580	4880808	239228
15	4907565	5049805	142240
15	5452492	5474330	21838
15	6705210	6818468	113258
15	7208463	7304325	95862
15	7317980	7335547	17567
15	7507678	7592890	85212
15	7682086	7700855	18769
18	2797554	2832090	34536
18	6702155	6728393	26238
26	5297874	5318387	20513
26	5517553	5550216	32663
30	183937	257768	73831
30	797497	844062	46565
30	3978481	4066243	87762
31	4041909	4078026	36117
32	4518067	4589712	71645
32	4616851	4625659	8808
32	4886114	4908718	22604
33	4163850	4200587	36737
36	1010120	1029957	19837
39	465841	688825	222984
39	1004596	1047034	42438
39	1207716	1236548	28832
39	2153726	2185052	31326
39	2245171	2269911	24740
39	2294746	2311616	16870
39	2323506	2340670	17164
40	1569517	1608679	39162
40	1870518	1891875	21357
43	3655049	3696771	41722
45	2785077	2828731	43654
51	1450783	1493272	42489
55	3423595	3500130	76535
57	46109	77869	31760
57	1615373	1638983	23610
64	55966	175700	119734
74	591451	600724	9273
77	2089974	2164351	74377
78	6039	59940	53901
82	2236206	2273645	37439
83	1379873	1425116	45243
84	2355047	2388352	33305
84	2399084	2517997	118913
88	819852	845401	25549
88	1194601	1316288	121687
88	1372483	1527476	154993
88	1732907	1868455	135548
88	1908746	1943289	34543
88	2418799	2435992	17193
91	129230	153938	24708
92	296055	342364	46309
93	1295671	1314656	18985
95	1001404	1044619	43215
97	298706	313982	15276
97	2158978	2172667	13689
97	2188270	2212097	23827
99	330458	339693	9235
99	355072	639642	284570
108	572788	738675	165887
108	814090	941030	126940
108	963573	1023559	59986
112	1966090	2014162	48072
113	1062779	1122847	60068
114	505730	526658	20928
114	1902474	2005892	103418
120	918534	962612	44078
121	1894243	1983613	89370
126	420889	439080	18191
143	1376555	1380941	4386
148	1458116	1644136	186020
148	1669247	1754172	84925
162	1227615	1263777	36162
164	0	113596	113596
164	196412	276496	80084
186	693304	711017	17713
190	805453	832175	26722
206	177202	290266	113064
229	252128	283116	30988
229	470627	578767	108140
304	0	70114	70114

Table S9

Candidate 'islands of speciation'. The maximum (**max**) d_{XY} , π_{diff}, and F_{ST} values for each putative 'island of speciation', together with the quantile in the corresponding distributions (**F_x**). Linkage group (**LG**) assignment for each scaffold (**sc**) as described in Methods - the **?** sign signifies that the scaffold could not be placed to a linkage group.



LG	**sc**	**start**	**end**	**max d_{XY}**	**F_x**	**max π_{diff}**	**F_x**	**max**	**F_x**
				$\times 10^{-3}$	**(max d_{XY})**	**$\times 10^{-4}$**	**(max π_{diff})**	**F_{ST}**	**(max F_{ST})**
LG5	15	1934238	1967068	1.53	95.44%	5.33	97.62%	0.52	99.97%
	15	2912637	2961336	1.35	92.98%	1.83	75.74%	0.43	99.93%
	15	4641580	4880808	1.68	96.54%	2.92	88.81%	0.56	99.99%
	15	4907565	5049805	2.09	98.09%	5.23	97.47%	0.45	99.95%
	15	6705210	6818468	2.07	98.06%	11.2	99.84%	0.59	99.99%
	15	7208463	7304325	1.60	96.01%	5.12	97.31%	0.62	100.00%
	15	7507678	7592890	1.75	96.98%	4.86	96.82%	0.63	100.00%
	18	6702155	6728393	1.48	94.86%	9.39	99.66%	0.38	99.84%
LG7	0	11529594	11540402	1.92	97.66%	2.53	85.30%	0.31	99.65%
	0	11994849	12015103	1.55	95.62%	2.92	88.87%	0.31	99.61%
	32	4518067	4589712	1.80	97.21%	3.17	90.64%	0.39	99.88%
	32	4886114	4908718	1.66	96.41%	6.35	98.68%	0.32	99.67%
	99	355072	639642	1.41	94.03%	7.50	99.25%	0.49	99.97%
LG12	14	3582492	3609841	1.36	93.22%	6.43	98.74%	0.32	99.70%
	14	3661853	3697260	1.45	94.47%	6.01	98.41%	0.51	99.97%
	43	3655049	3696771	1.62	96.21%	1.41	67.26%	0.34	99.77%
LG20	57	46109	77869	1.36	93.19%	6.82	98.99%	0.33	99.72%
	108	814090	941030	2.04	97.96%	4.08	94.84%	0.39	99.87%
	164	0	113596	1.27	91.55%	5.26	97.51%	0.30	99.58%
	164	196412	276496	1.60	96.02%	7.91	99.40%	0.31	99.62%
LG23	30	183937	257768	1.62	96.19%	8.06	99.44%	0.35	99.79%
	30	797497	844062	1.31	92.39%	3.92	94.25%	0.32	99.71%
	31	4041909	4078026	1.47	94.77%	2.39	83.83%	0.32	99.68%
	82	2236206	2273645	1.75	96.99%	7.26	99.17%	0.36	99.80%
	88	819852	845401	1.27	91.42%	4.70	96.51%	0.34	99.76%
	88	1194601	1316288	1.50	95.08%	11.3	99.89%	0.41	99.90%
	88	1372483	1527476	1.86	97.48%	11.3	99.88%	0.46	99.95%
	88	1732907	1868455	1.38	93.55%	10.1	99.76%	0.35	99.79%
	95	1001404	1044619	1.35	93.01%	6.99	99.08%	0.33	99.71%
LG8	51	1450783	1493272	1.53	95.44%	3.93	94.31%	0.31	99.67%
	113	1062779	1122847	3.11	99.27%	5.08	97.25%	0.43	99.93%
	190	805453	832175	1.21	90.13%	1.08	58.40%	0.33	99.73%
LG3	126	420889	439080	3.78	99.53%	26.6	99.99%	0.32	99.69%
	186	693304	711017	1.38	93.60%	3.08	90.02%	0.30	99.60%
LG19	120	918534	962612	1.99	97.86%	4.35	95.64%	0.36	99.81%
	162	1227615	1263777	1.47	94.78%	1.49	69.10%	0.31	99.63%
?	39	465841	688825	2.37	98.59%	5.94	98.35%	0.47	99.96%
	39	2323506	2340670	1.22	90.40%	1.86	76.19%	0.31	99.66%
?	148	1458116	1644136	2.48	98.75%	10.3	99.74%	0.50	99.97%
	148	1669247	1754172	2.28	98.45%	6.05	98.45%	0.33	99.74%
LG18	6	2399603	2417150	1.46	94.65%	9.49	99.68%	0.30	99.61%
LG2	11	5426321	5452278	1.42	94.17%	2.97	89.21%	0.33	99.72%
LG13	26	5297874	5318387	1.59	95.97%	2.40	83.92%	0.30	99.59%
LG4	55	3423595	3500130	1.61	96.12%	4.76	96.63%	0.42	99.91%
LG11	64	55966	175700	1.42	94.08%	8.53	99.53%	0.34	99.75%
LG15	78	6039	59940	1.49	94.99%	2.28	82.59%	0.38	99.85%
LG14	84	2399084	2517997	1.40	93.81%	11.0	99.79%	0.38	99.85%
LG6	97	2188270	2212097	1.31	92.28%	2.25	82.16%	0.32	99.68%
LG9	229	470627	578767	1.87	97.53%	7.42	99.23%	0.39	99.86%
?	45	2785077	2828731	1.74	96.94%	3.24	91.09%	0.31	99.61%
?	91	129230	153938	1.62	96.20%	0.87	51.23%	0.34	99.75%
?	112	1966090	2014162	1.29	91.87%	7.81	99.38%	0.32	99.71%
?	114	1902474	2005892	2.10	98.13%	11.2	99.79%	0.32	99.70%
?	206	177202	290266	1.46	94.63%	8.66	99.56%	0.38	99.84%
?	304	0	70114	1.14	100.00%	18.9	99.98%	0.41	99.91%

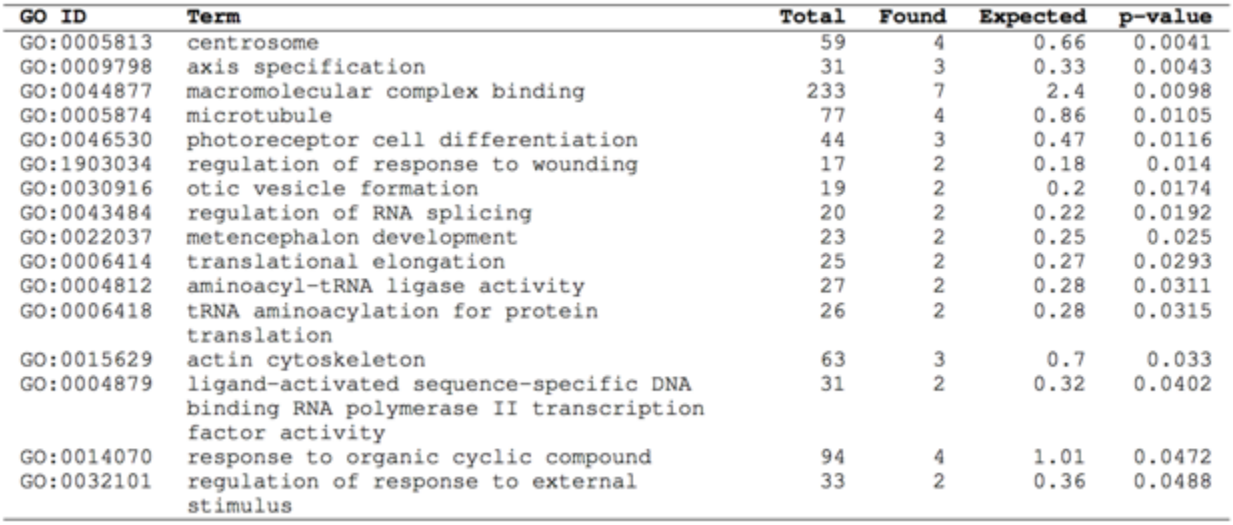
Table S10

Genotyped variants used for mate-choice trials and F_{ST} values observed in the reference sample of 18 benthic and 16 littoral males. F_{ST} values are based only on the genotyped 18 Massoko benthic and 16 littoral males (the whole-genome sequenced individuals and other individuals used in the mate-choice trial are not included).

Variant coordinates	F_{ST}
scaffold_1:4365113	0.466
scaffold_1:4365291	0.466
scaffold_6:7294300	0.405
scaffold_7:3305104	0.291
scaffold_7:3305216	0.291
scaffold_7:3313226	0.323
scaffold_7:3318131	0.262
scaffold_7:3318579	0.262
scaffold_12:3793589	0.416
scaffold_12:3793994	0.416
scaffold_14:3663857	0.424
scaffold_14:3669941	0.424
scaffold_14:4168852	0.232
scaffold_15:2959443	0.396
scaffold_15:2962256	0.135
scaffold_15:5455316	0.22
scaffold_15:5458471	0.174
scaffold_15:7238850	0.659
scaffold_15:7251797	0.701
scaffold_15:7252309	0.657
scaffold_15:7254754	0.659
scaffold_15:7269475	0.659
scaffold_15:7269758	0.701
scaffold_18:4359768	0.082
scaffold_18:4362575	0.038
scaffold_19:377749	0.049
scaffold_26:1611369	0.038
scaffold_29:3346390	0.011
scaffold_30:4055115	0.458
scaffold_30:6447512	0.194
scaffold_30:6448182	0.307
scaffold_30:6452026	0.281
scaffold_31:426382	0
scaffold_31:1867496	0.104
scaffold_34:2235172	0.112
scaffold_34:2250784	0.152
scaffold_34:2802787	0.026
scaffold_35:1693366	0.071
scaffold_38:642672	0.388
scaffold_40:1588650	0.292
scaffold_40:1588728	0.292
scaffold_40:1621279	0.273
scaffold_40:1882810	0.206
scaffold_49:3025847	0.503
scaffold_55:2299953	0.025
scaffold_55:3423696	0.419
scaffold_55:3424572	0.456
scaffold_55:3435034	0.382
scaffold_55:3435507	0.382
scaffold_55:3451516	0.303
scaffold_55:3453112	0.303
scaffold_55:3480538	0.345
scaffold_55:3483480	0.378
scaffold_67:1282346	0.171
scaffold_67:1284312	0.171
scaffold_67:1288981	0.219
scaffold_82:2709101	0.076
scaffold_82:2724117	0.076
scaffold_82:2731927	0.098
scaffold_87:112	0.159
scaffold_87:5003	0.159
scaffold_87:50289	0.098
scaffold_87:51344	0.072
scaffold_88:1185176	0.198
scaffold_88:1198991	0.496
scaffold_88:1199168	0.382
scaffold_88:1213550	0.496
scaffold_88:1213711	0.462
scaffold_88:1312010	0.616
scaffold_88:1312055	0.616
scaffold_88:1441936	0.71
scaffold_88:1484291	0.71
scaffold_88:1488398	0.71
scaffold_88:1539583	0.452
scaffold_88:1647616	0.387
scaffold_88:1654621	0.387
scaffold_88:1675293	0.387
scaffold_88:1781770	0.71
scaffold_88:1786645	0.71
scaffold_88:1786888	0.677
scaffold_88:1825810	0.71
scaffold_88:1825839	0.742
scaffold_88:1886989	0.665
scaffold_88:1923134	0.71
scaffold_88:1940222	0.581
scaffold_88:1940476	0.677
scaffold_88:1942123	0.581
scaffold_88:2222087	0.243
scaffold_88:2222122	0.243
scaffold_88:2429606	0.665
scaffold_88:2470678	0.345
scaffold_88:2473383	0.345
scaffold_88:2487048	0.135
scaffold_91:11230	0.101
scaffold_91:54791	0.159
scaffold_91:55547	0.159
scaffold_91:117365	0.092
scaffold_91:528794	0
scaffold_91:530091	0.008
scaffold_97:188537	0.145
scaffold_97:193146	0.118
scaffold_97:193356	0.19
scaffold_97:2055581	0.295
scaffold_126:32880	0.279
scaffold_126:38797	0.249
scaffold_126:1275145	0.264
scaffold_126:1284768	0.377
scaffold_146:1672851	0
scaffold_155:1053156	0.232
scaffold_217:517341	0.419
scaffold_241:14395	0.03
scaffold_241:16987	0.162
scaffold_241:32682	0.165
scaffold_259:243418	0.355
scaffold_259:249905	0.387
scaffold_316:212810	0.456
scaffold_316:213151	0.456

Table S11

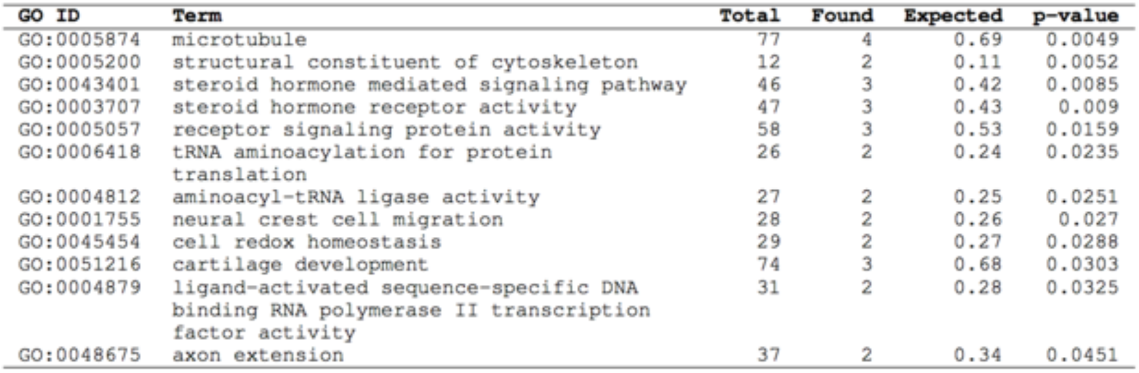
Gene Ontology (GO) enrichment terms in candidate ‘islands of speciation’ ±50kb.



GO ID	Term	Total	Found	Expected	p-value
GO:0005813	centrosome	59	4	0.66	0.0041
GO:0009798	axis specification	31	3	0.33	0.0043
GO:0044877	macromolecular complex binding	233	7	2.4	0.0098
GO:0005874	microtubule	77	4	0.86	0.0105
GO:0046530	photoreceptor cell differentiation	44	3	0.47	0.0116
GO:1903034	regulation of response to wounding	17	2	0.18	0.014
GO:0030916	otic vesicle formation	19	2	0.2	0.0174
GO:0043484	regulation of RNA splicing	20	2	0.22	0.0192
GO:0022037	metencephalon development	23	2	0.25	0.025
GO:0006414	translational elongation	25	2	0.27	0.0293
GO:0004812	aminoacyl-tRNA ligase activity	27	2	0.28	0.0311
GO:0006418	tRNA aminoacylation for protein translation	26	2	0.28	0.0315
GO:0015629	actin cytoskeleton	63	3	0.7	0.033
GO:0004879	ligand-activated sequence-specific DNA binding RNA polymerase II transcription factor activity	31	2	0.32	0.0402
GO:0014070	response to organic cyclic compound	94	4	1.01	0.0472
GO:0032101	regulation of response to external stimulus	33	2	0.36	0.0488

Table S12

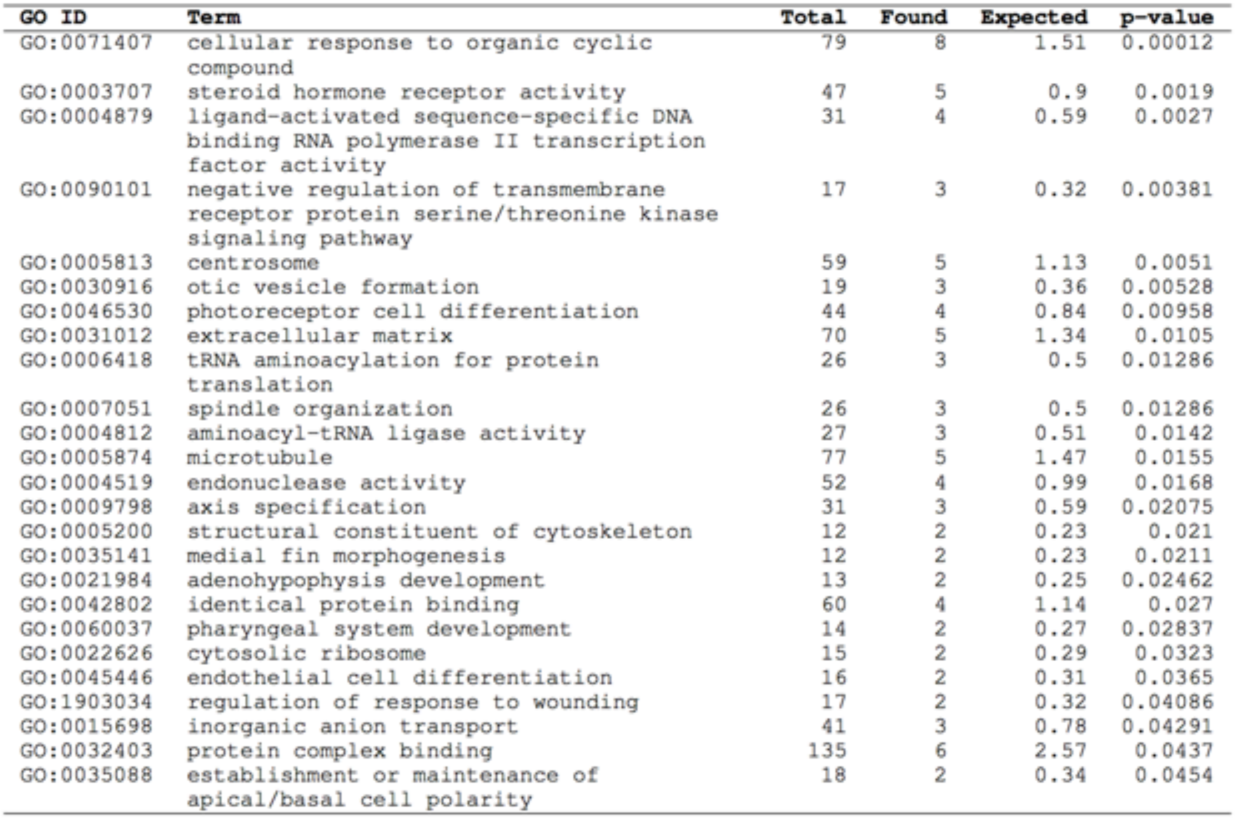
GO terms significantly enriched in all HDRs ±10kb.



GO ID	Term	Total	Found	Expected	p-value
GO:0005874	microtubule	77	4	0.69	0.0049
GO:0005200	structural constituent of cytoskeleton	12	2	0.11	0.0052
GO:0043401	steroid hormone mediated signaling pathway	46	3	0.42	0.0085
GO:0003707	steroid hormone receptor activity	47	3	0.43	0.009
GO:0005057	receptor signaling protein activity	58	3	0.53	0.0159
GO:0006418	tRNA aminoacylation for protein translation	26	2	0.24	0.0235
GO:0004812	aminoacyl-tRNA ligase activity	27	2	0.25	0.0251
GO:0001755	neural crest cell migration	28	2	0.26	0.027
GO:0045454	cell redox homeostasis	29	2	0.27	0.0288
GO:0051216	cartilage development	74	3	0.68	0.0303
GO:0004879	ligand-activated sequence-specific DNA binding RNA polymerase II transcription factor activity	31	2	0.28	0.0325
GO:0048675	axon extension	37	2	0.34	0.0451

Table S13

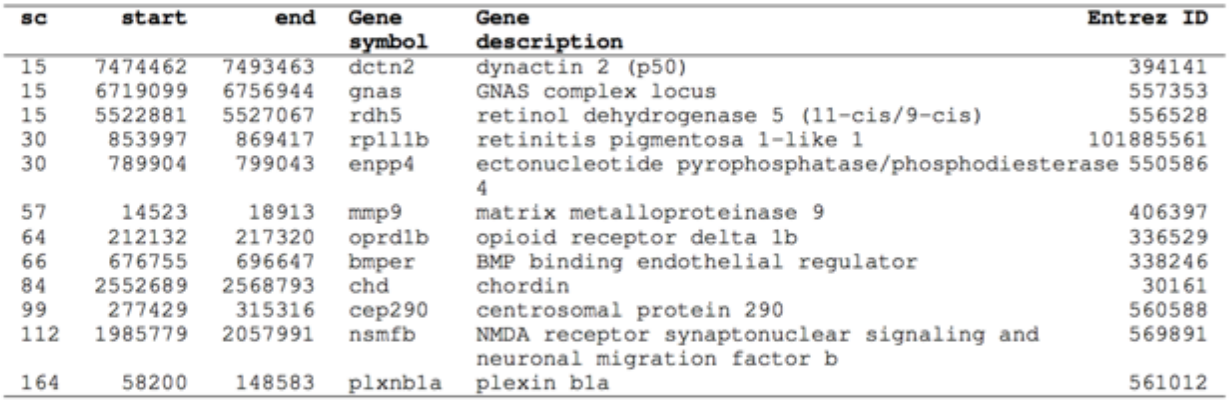
GO terms significantly enriched in all HDRs ±50kb.



GO ID	Term	Total	Found	Expected	p-value
GO:0071407	cellular response to organic cyclic compound	79	8	1.51	0.00012
GO:0003707	steroid hormone receptor activity	47	5	0.9	0.0019
GO:0004879	ligand-activated sequence-specific DNA binding RNA polymerase II transcription factor activity	31	4	0.59	0.0027
GO:0090101	negative regulation of transmembrane receptor protein serine/threonine kinase signaling pathway	17	3	0.32	0.00381
GO:0005813	centrosome	59	5	1.13	0.0051
GO:0030916	otic vesicle formation	19	3	0.36	0.00528
GO:0046530	photoreceptor cell differentiation	44	4	0.84	0.00958
GO:0031012	extracellular matrix	70	5	1.34	0.0105
GO:0006418	tRNA aminoacylation for protein translation	26	3	0.5	0.01286
GO:0007051	spindle organization	26	3	0.5	0.01286
GO:0004812	aminoacyl-tRNA ligase activity	27	3	0.51	0.0142
GO:0005874	microtubule	77	5	1.47	0.0155
GO:0004519	endonuclease activity	52	4	0.99	0.0168
GO:0009798	axis specification	31	3	0.59	0.02075
GO:0005200	structural constituent of cytoskeleton	12	2	0.23	0.021
GO:0035141	medial fin morphogenesis	12	2	0.23	0.0211
GO:0021984	adenohypophysis development	13	2	0.25	0.02462
GO:0042802	identical protein binding	60	4	1.14	0.027
GO:0060037	pharyngeal system development	14	2	0.27	0.02837
GO:0022626	cytosolic ribosome	15	2	0.29	0.0323
GO:0045446	endothelial cell differentiation	16	2	0.31	0.0365
GO:1903034	regulation of response to wounding	17	2	0.32	0.04086
GO:0015698	inorganic anion transport	41	3	0.78	0.04291
GO:0032403	protein complex binding	135	6	2.57	0.0437
GO:0035088	establishment or maintenance of apical/basal cell polarity	18	2	0.34	0.0454

Table S14

Genes contributing to GO enriched terms in sensory perception. The genomic location is given in *M. zebra* assembly coordinates (**sc**-scaffold). Gene symbol, description and Entrez ID are given for zebrafish orthologs.



sc	start	end	Gene symbol	Gene description	Entrez ID
15	7474462	7493463	dctn2	dynactin 2 (p50)	394141
15	6719099	6756944	gnas	GNAS complex locus	557353
15	5522881	5527067	rdh5	retinol dehydrogenase 5 (11-cis/9-cis)	556528
30	853997	869417	rp1l1b	retinitis pigmentosa 1-like 1	101885561
30	789904	799043	enpp4	ectonucleotide pyrophosphatase/phosphodiesterase 4	550586
57	14523	18913	mmp9	matrix metalloproteinase 9	406397
64	212132	217320	oprd1b	opioid receptor delta 1b	336529
66	676755	696647	bmper	BMP binding endothelial regulator	338246
84	2552689	2568793	chd	chordin	30161
99	277429	315316	cep290	centrosomal protein 290	560588
112	1985779	2057991	nsmfb	NMDA receptor synaptonuclear signaling and neuronal migration factor b	569891
164	58200	148583	plxnb1a	plexin b1a	561012

Table S15

Shimodaira-Hasegawa tests (*43*) of single-lake monophyly constraints against an unconstrained topology using RAxML 8.0.22 (*41*). The test suggests we can reject the hypothesis of monophyly for lake Itamba (P<0.05). Monophyly of each of the other lakes cannot be rejected.

Monophyly constraint	Likelihood	D(LH)	SD	P(worse)
None	-53503.81			
Massoko	-53518.76	-14.95	29.05	> 0.05
Itende	-53515.48	-11.67	28.66	> 0.05
Ikapu	-53515.48	-11.67	28.66	> 0.05
Itamba	-53533.66	-29.84	14.92	< 0.05
Ilamba	-53515.89	-12.08	28.66	> 0.05
Kingiri	-53515.48	-11.67	28.66	> 0.05

Table S16

Individual BioSample accessions for whole genome sequencing data.

Samples	BioSample Accessions
A. calliptera	SAMEA1904323, SAMEA1904326-SAMEA1904328, SAMEA1920090, SAMEA1920092, SAMEA2661381-SAMEA2661387, SAMEA2661389-SAMEA2661391
Massoko benthic	SAMEA1877404, SAMEA1877436, SAMEA1877511, SAMEA1877425, SAMEA1877494, SAMEA1877464, SAMEA2661333-SAMEA2661339, SAMEA2661341-SAMEA2661347, SAMEA2661349-SAMEA2661355, SAMEA2661357-SAMEA2661359, SAMEA2661362, SAMEA2661363, SAMEA2661365-SAMEA2661370
Massoko littoral	SAMEA1877400, SAMEA1877402, SAMEA1877407, SAMEA1877442, SAMEA1877447, SAMEA1877507, SAMEA2661297-SAMEA2661301, SAMEA2661303,SAMEA2661304, SAMEA2661306-SAMEA2661310, SAMEA2661313, SAMEA2661316, SAMEA2661318-SAMEA2661322, SAMEA2661324-SAMEA2661328, SAMEA2661330, SAMEA2661331
Massoko small	SAMEA2661371, SAMEA2661373-SAMEA2661380, SAMEA2661392-SAMEA2661400, SAMEA2661402-SAMEA2661414
Lake Itamba	SAMEA2661415, SAMEA2661417-SAMEA2661440, SAMEA2661442-SAMEA2661446

Movie S1

Lake Massoko *Astatotilapia* ecomorphs. (A) Lake Massoko panorama. **(B-E)** Littoral and benthic males recorded in their natural habitats. Two different littoral males are shown at depths of ~1m and ~4m. Two different benthic males are shown at depths between 20-25m with light provided by a hand-held torch.

Movie S2

Mate-choice experiments. The video illustrates aspects of the experimental setup, including: **(A)** females passing through the 'partial partition' grid and **(B)** large males confined to their territories. **(C-D)** Examples of females being courted by benthic and littoral males during the trials.