nature COMMUNICATIONS

ARTICLE

Received 23 Sep 2015 | Accepted 28 Jul 2016 | Published 6 Sep 2016

DOI: 10.1038/ncomms12736 OPEN

Genetic linkage of distinct adaptive traits in sympatrically speciating crater lake cichlid fish

Carmelo Fruciano[1,2,*], Paolo Franchini[1,*], Viera Kovacova[1,3], Kathryn R. Elmer[1,4], Frederico Henning[1] & Axel Meyer[1]

Our understanding of how biological diversity arises is limited, especially in the case of speciation in the face of gene flow. Here we investigate the genomic basis of adaptive traits, focusing on a sympatrically diverging species pair of crater lake cichlid fishes. We identify the main quantitative trait loci (QTL) for two eco-morphological traits: body shape and pharyngeal jaw morphology. These traits diverge in parallel between benthic and limnetic species in the repeated adaptive radiations of this and other fish lineages. Remarkably, a single chromosomal region contains the highest effect size QTL for both traits. Transcriptomic data show that the QTL regions contain genes putatively under selection. Independent population genomic data corroborate QTL regions as areas of high differentiation between the sympatric sister species. Our results provide empirical support for current theoretical models that emphasize the importance of genetic linkage and pleiotropy in facilitating rapid divergence in sympatry.

[1] Lehrstuhl für Zoologie and Evolutionsbiologie, Department of Biology, University of Konstanz, Universitätsstrasse 10, 78457 Konstanz, Germany. [2] School of Earth, Environmental and Biological Sciences, Queensland University of Technology, Brisbane, Queensland 4000, Australia. [3] Department for Plant Developmental Genetics, Institute of Biophysics, Academy of Sciences Czech Republic, Královopolská 135, 612 65 Brno, Czech Republic. [4] Institute of Biodiversity, Animal Health and Comparative Medicine, College of Medical, Veterinary and Life Sciences University of Glasgow, Glasgow G12 8QQ, UK. * These authors contributed equally to this work. Correspondence and requests for materials should be addressed to A.M. (email: axel.meyer@uni-konstanz.de).

Sympatric speciation—the process by which new species arise in a geographic setting without barriers to gene flow—has been hotly debated over the past 50 years[1,2]. It has been controversial because of the restrictive genetic and environmental conditions probably needed for divergent selection to overcome the homogenizing effects of gene flow and produce different species in the absence of extrinsic barriers[2–4]. To date, only a small number of empirical examples are accepted to have fulfilled the conditions for speciation with gene flow: divergent host races in herbivorous insects[5], divergence in palms and other plants promoted by diversity in soil on remote oceanic islands[6] and trophically polymorphic crater lake cichlid fishes[7,8]. Consequently, the research focus has shifted to understanding the conditions that permit or promote sympatric speciation[2,4,9]. Most models of speciation, whether mathematical or verbal, require a strong role of close linkage[2,10] of the genetic loci that underlie the diverging phenotypic traits. These include 'divergence hitchhiking' models, in which a decrease of gene flow between populations in genomic regions surrounding loci under divergent selection can generate larger regions of differentiation between the diverging genomes[11]. Empirical tests of theoretical models on the few substantiated cases of sympatric speciation are, however, still scarce (but see refs 5,12).

Nicaraguan crater lake cichlid fishes are one of those few well-substantiated instances of sympatric speciation[3,7,8]. The chain of crater lakes in Nicaragua has been independently colonized from the large and shallow Nicaraguan great lakes. Crater Lake Apoyo is maximally *ca.* 22,000 years old, small, deep and characterized by clear water[13]. A small and monophyletic adaptive radiation of six endemic cichlid species (*Amphilophus* spp. complex, or Midas cichlids) has formed rapidly[8,13] and sympatrically into open water (limnetic) and bottom-dwelling (benthic) species[7]. These species differ in body shape (limnetics are more elongate)[13,14] and trophic ecology[8], have different gut bacterial communities[15] and differ in the morphology of their lower pharyngeal jaws[7] (modified gill arches that form a functional second jaw and are used to crush hard food; these are more robust in benthic forms; Fig. 1). Body shape has important ecological consequences and a genetic basis, as the difference between species is retained when fish are grown under common laboratory conditions[14]. The adaptive significance of variation in body shape has been shown directly in sticklebacks[16] and perch[17], where deep bodied fish perform better in benthic environments and more elongated fish perform better in the open water. Biomechanical studies of fish locomotion also suggest that a deeper body performs better when maneuvering in the more structurally complex benthic zone[18]. The adaptive significance of variation in pharyngeal jaw morphology in Midas cichlids has been evidenced experimentally. Indeed, the more robust pharyngeal jaws typical of the benthic forms perform better when processing hard food as compared with softer food items, whereas the opposite is true for the more gracile pharyngeal jaws of limnetic species[19]. The sympatric speciation and the well-studied traits involved in ecological divergence make the Midas cichlid species flock an ideal system to clarify the genetic basis of ecologically relevant traits and how selective pressures can translate into important genomic differences in the face of ongoing gene flow.

Theory predicts that linkage and/or pleiotropy might increase the likelihood or even facilitate sympatric speciation more generally, but especially as it applies to Midas cichlids[20–22]. Here we show that the quantitative trait locus (QTL) of largest effect for body shape and pharyngeal jaw shape overlap in a single linkage group (LG), making these two key ecological traits genetically non-independent. We also identify the co-localization of genes under selection and QTL for these traits. Our results, then, provide empirical support for current theoretical models of speciation with gene flow.

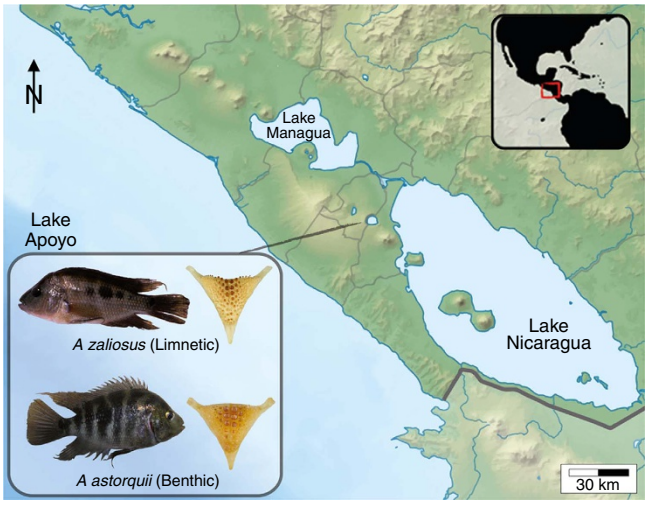

Figure 1 | Nicaraguan lakes and benthic/limnetic Midas cichlids. The lake indicated is the crater lake Apoyo. In the inset, pictures of representative specimens of *A. astorquii* and *A. zaliosus*, and of typical lower pharyngeal jaws of these species.

Results

Mapping of QTL. To identify the genomic regions underlying ecologically relevant morphology, we performed QTL mapping on an interspecies genetic cross of benthic *Amphilophus astorquii* and limnetic *Amphilophus zaliosus*. From high-coverage sequencing of double-digest restriction site-associated DNA (ddRAD), we constructed a genetic map of 495 single-nucleotide polymorphisms (SNPs) resolving 24 LGs with an average marker spacing of 2.65 cM. This map was then used for multivariate QTL mapping of body shape and pharyngeal jaw size and shape. We found highly supported QTL for both shape traits (five QTL for body shape and three for pharyngeal jaw shape; see Supplementary Table 1 and Supplementary Figs 1 and 2 for location, confidence interval, effect size and predicted shape change of each QTL). We did not find any significant QTL for lower pharyngeal jaw size. Notably, the QTL of strongest effect for both body shape and pharyngeal jaw shape co-located on LG 3 and their confidence intervals overlap (Bayesian credibility intervals 18–28 cM for body, 6–24 cM for lower pharyngeal jaw shape; Fig. 2). This genetic non-independence is further supported by the analysis showing significant covariation (Escoufier $RV = 0.033$, $P = 0.044$) in body and pharyngeal jaw shape in the QTL mapping population of F_2 individuals. When mapping covariation of body and pharyngeal jaw shape, a single highly supported QTL on LG 3 was found (Bayesian credibility interval 16–31 cM; Fig. 2). Finally, a statistical test based on permutations shows (overlapPermTest function of the regioneR package[23], overlap = 1, $P = 0.044$) that the overlap of the QTL regions for body and pharyngeal jaw shape is higher than expected by chance. These results suggest a pleiotropic effect or a close linkage between the genetic loci underlying quantitative variation in these two ecomorphological traits.

Co-localization of genes under selection and QTL regions. To identify what genes are potentially contributing to the benthic–limnetic divergence between species in Lake Apoyo by responding to positive selection, and whether they co-localize with QTL

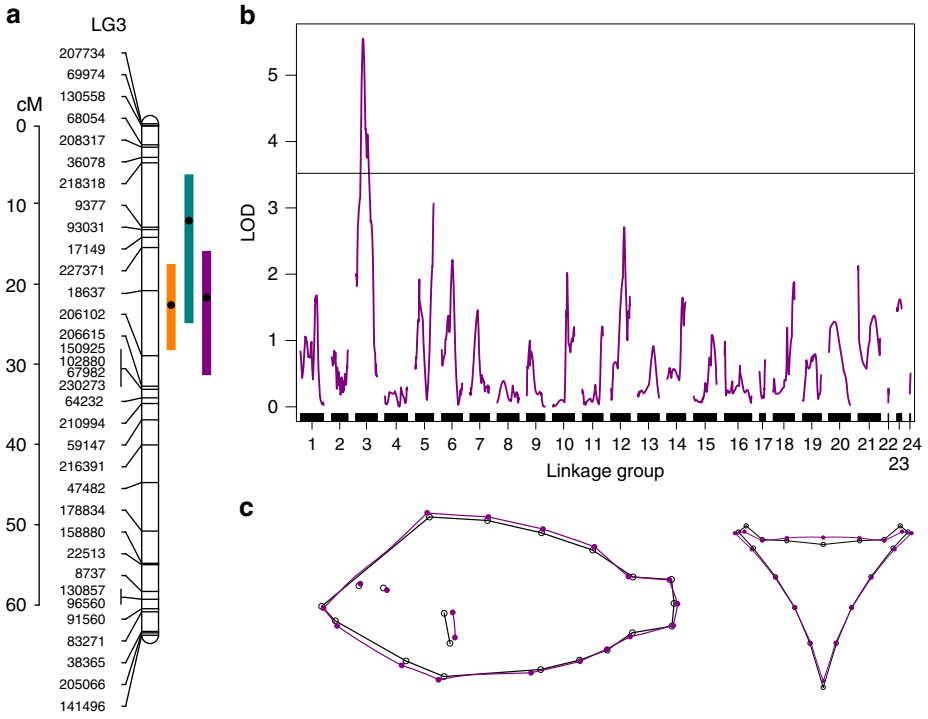

Figure 2 | QTL mapping. (**a**) Bayesian credibility intervals for QTL in LG 3: orange bars for body shape; green bars for pharyngeal jaw shape; violet bar for covariation of body shape and pharyngeal jaw shape. (**b**) LOD scores at each position in each LG obtained by mapping PLS scores (covariation). The horizontal line identifies the genome-wide significance threshold obtained through permutations (at LOD = 3.52). (**c**) Covariation accounted for by the only QTL (on LG 3) with LOD score higher than the genome-wide significance threshold.

regions, we sequenced a Midas reference transcriptome combining the reads of five closely related species of the *Amphilophus* complex by Illumina next-generation sequencing. After merging a reference-guided and a *de novo* assembly, and retaining the transcript with the highest similarity score among those matching the same Nile tilapia protein, we obtained a Midas reference transcriptome of 15,348 sequences (N50 value: 5,067). After mapping the reads of our two focal species (*A. astorquii* and *A. zaliosus*) and extracting the species consensus (see Methods), we obtained a total of 11,103 genes orthologous between species, of which 71 showed signatures of positive selection (dN/dS >1; Supplementary Data 1). These genes had no significant over-representation of functional categories when compared against the full transcriptomic data set of orthologous genes, as revealed by enrichment analyses (Fisher's exact test; $P>0.05$). Mapping these 71 genes showing signatures of positive selection and all the linkage map markers on the Midas cichlid draft genome[8], we found that three genes showing positive selection were associated with two QTL regions for the mapped phenotypic traits. One of these genes (*apolipoprotein eb-like*) co-located in a QTL region (in LG9) for body shape, whereas two genes (*cell cycle control protein 50a-like* and *sodium-coupled neutral amino acid transporter 4-like isoform x6*) clustered with the QTL for pharyngeal jaw shape on LG12 (see Supplementary Data 1 for details).

Genetic divergence in QTL regions. To test whether the QTL regions co-localize with regions of genetic divergence between species in natural populations, we used pooled population genomic data of wild-caught individuals[8]. Although the average genome-wide differentiation between species was relatively low (average $F_{ST}=0.083$ based on a 500 bp sliding window; Supplementary Data 2), the regions of the Midas cichlid genome unequivocally attributable to QTL regions (see Methods) were more differentiated between species (average $F_{ST}=0.097$) than were the genomic regions containing non-QTL linkage map markers (average $F_{ST}=0.081$, similar to the genome-wide F_{ST} computed using all data). Three genomic regions containing markers in the QTL intervals also show significant differences between species in allele frequencies (Fisher's exact test $P<0.01$ after correcting for multiple tests; Fig. 3 and Supplementary Data 3).

Discussion

We have identified significant signals of positive selection and genomic co-localization of QTL underlying two major ecologically relevant traits in sympatrically speciating sister species of cichlids. Specifically, in cichlid fish from Lake Apoyo we identified five QTL for body shape and three for pharyngeal jaw shape. However, we did not identify any QTL for pharyngeal jaw size. This, combined with the fact that there is no significant difference in pharyngeal jaw size between the two species when kept under captive conditions, suggests that the variation in pharyngeal jaw size between these species is perhaps more affected by the environment. Indeed, although the existence of pharyngeal jaws and their diversification have been implicated in playing a causal role in the spectacular diversification of cichlids, these structures are well known to exhibit phenotypically plastic variation depending on the hardness of consumed food items[24]. The QTL for body shape show diverse effects on this trait (Supplementary Fig. 1c). It is remarkable that the QTL of strongest effect for body and pharyngeal jaw shape, overlapping in LG3, produce, respectively, a deepening of the middle of the body and an enlargement of the posterior margin of the pharyngeal jaw. These shape changes match the variation between species previously described in natural populations[7,13]

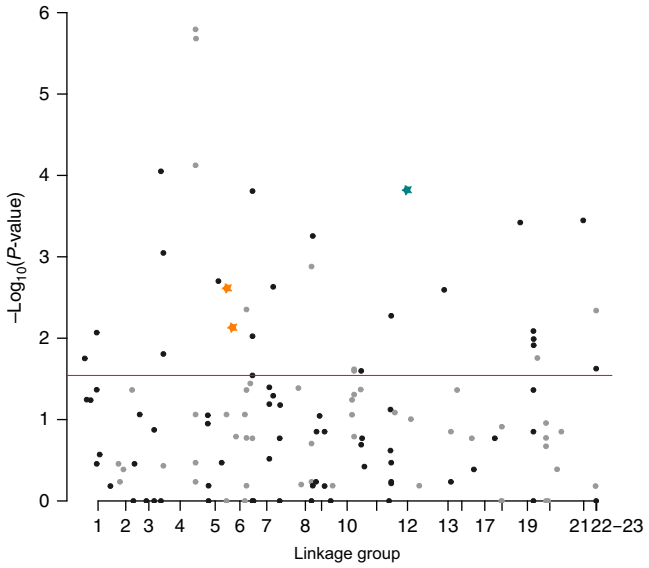

Figure 3 | Population genomics. Manhattan plot of genome-wide differentiation between sympatric Midas species (significance levels for the Fisher test), including regions that co-localize with QTL. Only genomic windows containing RAD markers in the linkage map are plotted. The horizontal line represents a significance threshold obtained from the multiple test correction procedure ($-\mathrm{Log}_{10}$ (*P*-value) = 1.49). Coloured stars represent genomic windows significantly differentiated between species and located in QTL regions (orange for body shape, green for pharyngeal jaw shape). SNPs not located in QTL regions are represented in grey or black, alternating these two colors so that the LGs are distinguishable.

and have a clear functional relevance. Indeed, a deeper body is advantageous for swimming in a benthic environment[18], where an enlarged pharyngeal jaw is beneficial[19] because of the higher abundance of hard prey. The genomic regions containing QTL for body and pharyngeal jaw shape also contain expressed genes that show signals of response to positive selection and their function has been linked to growth and cell cycle control[25]. For this reason, these three genes are promising candidates that might contribute to the observed morphological diversification between the two focal species. However, experimental validation is necessary to clarify their causal role.

Further, the regions containing QTL showed significantly different allele frequencies between species, thus giving a clear genomic signal of divergent natural selection acting on these traits or at tightly linked genes.

This finding is extremely important as current models of speciation emphasize the importance of tight linkage and pleiotropy in rapid speciation[26], yet, so far, there is a lack of empirical examples supporting them. Recent simulation studies[21] show that the arrangement of genes in genomes is important to facilitate divergence. In particular, not only is the statistical association among large numbers of genes important for rapid divergence, but so also is divergent selection on persistent allelic combinations among these genes[22]. These genomic associations are important factors, especially in sympatry, and can alleviate the need for assortative mating[21]. The exact role of tight linkage and pleiotropy in rapid speciation is still debated—as they can both conceivably favour and hinder divergence[26]. In the presence of very strong divergent selection, linkage might not contribute substantially to the establishment of divergently selected alleles (that is, they would diverge anyway whether linked or not, whereas with milder divergent selection linkage might be more important in favouring divergence). Our data, however, does not support this scenario in Midas cichlids. In fact, we would expect that very strong divergent selection on a few loci would result in a strong divergence in surrounding regions, which would be reflected in much higher F_{ST} estimates in QTL regions compared with the rest of the genome. In our case, instead, F_{ST} estimates in the QTL regions are just slightly higher than in the rest of the genome, suggesting not very strong divergent selection, a condition in which linkage could play a substantial role in promoting sympatric divergence.

The most recent theoretical predictions based on simulations[21] suggest that pleiotropy or tight linkage are particularly—but not exclusively—important in favouring divergence in conditions of many genes under selection, genes of small effect and high gene flow. The number of genes under selection is important, because genomes tend to become more consolidated if many loci are contributing to divergence[21]. Clearly, in our case an accurate estimate of the number of genes underlying adaptive traits and under divergent selection is not possible due to empirical detection limits, such as the difficulty of detecting small-effect QTL. However, our empirical results generally conform to these recent theoretical expectations, as we find 71 candidate genes under selection and eight QTL regions for two traits. It is also likely to be that we failed to detect a number of smaller effect QTL, and that some of these might be genetically non-independent across the two traits.

High gene flow is supported in Apoyo Midas cichlids by the low average F_{ST}, which is concordant with previous estimates on the same system based on different genetic markers[7,27]. Similar F_{ST} estimates have been documented in situations of divergence with gene flow in sticklebacks[28]. Low values of F_{ST} are probably also due to recent divergence. Indeed, the values of F_{ST} we observe are of similar magnitude to the ones observed in populations of another cichlid fish, *Archocentrus centrarchus*[29], which recently diverged in allopatry in Nicaraguan lakes.

The last condition under which pleiotropy or linkage are expected to substantially favour divergence—the condition of genes of small effect—seems also to be satisfied in our study system. Effect sizes are not easily comparable across studies, as the inherent bias in their estimation[30] might vary in extent depending on how a trait is defined and how it is mapped. However, other QTL studies on body shape have reported much higher effect sizes (typically around 10% in at least one QTL, sometimes as high as 20–30%) in a range of fish species[31,32]. In our study, instead, the QTL of largest genetic additive effect for any of the two traits accounts for a mere 4.08% of trait variance. Then, the effect sizes for QTL of body and pharyngeal jaw shape that we found in this study seem also to satisfy the condition of genes of small effect.

Thus, by largely matching theoretical predictions, our empirical results overall suggest that pleiotropy or tight linkage of QTL for different traits can facilitate rapid sympatric divergence.

Methods

Data sets used. For QTL mapping, a total of 305 F_2 individuals from an *A. astorquii* × *A. zaliosus* cross were used in the present study. Briefly, a wild-caught female *A. astorquii* was crossed with a wild-caught male *A. zaliosus* and the eggs were removed from the parental tank once spawned. On maturity (*ca.* 1 year of age), we randomly chose one pair (F_1) as it formed, isolated them into a different tank and allowed them to breed. This F_1 pair produced the F_2 individuals used here for QTL mapping. All fish were photographed in a standardized manner for morphometric analyses of body shape at 18 months of age. Further details on the cross are provided in ref. 14, where we also show that the two parental species retain differences in body shape even when raised under the same laboratory conditions. We performed a preliminary exploratory analysis of pharyngeal jaw size and shape on a small set (five *A. astorquii* and ten *A. zaliosus*) of wild-caught lab-reared individuals of each parental species. These showed nonsignificant difference in pharyngeal jaw size (analysis of covariance using body centroid size as

a covariate for allometric correction; $P=0.32$) but significant difference in average pharyngeal jaw shape (Procrustes distance 0.042; $P=0.0014$ based on 10,000 permutations). The individuals used here include the ones analysed in ref. 14 (according to German law on animal welfare and specifically approved by the Regierungspräsdium Freiburg, Abteilung Landwirtschaft, Ländlicher Raum, Veterinär- und Lebensmittelwesen; approval G-11/ 73 35-9185.81), complemented by new specimens, which had not been sequenced before. Individuals of the F_2 mapping population were tagged and later killed to dissect the lower pharyngeal jaw ($n=265$ due to part of the mapping population losing their transponder tag) at two time points. At this stage, we also took a second picture of their body to compute body centroid size and to analyse covariation between body and pharyngeal jaw shapes. To quantify the morphology of body and pharyngeal jaws, we used geometric morphometrics and analysed the set of points depicted in Supplementary Fig. 3. As in our previous study[14], configurations of points were aligned through a generalized Procrustes analysis with sliding of semi-landmarks[33] and allometric variation was removed from the data using, for both body and pharyngeal jaw shape, residuals of regression on body centroid size. In the case of body shape, before this regression, we also carried out a procedure for removal of body arching[34–36]. For pharyngeal jaws, we performed the analysis of shape only on the symmetric component[37], as this is the ecologically relevant component of shape variation that distinguish the two species (that is, enlarged pharyngeal jaws have been described in benthic Midas cichlids, as opposed to the more gracile pharyngeal jaw of the open water species). We used, as measures of pharyngeal jaw size, pharyngeal jaw centroid size and pharyngeal jaw weight. Each of these measures of pharyngeal jaw size was regressed on body centroid size and the residuals were used in subsequent analyses so as to account for allometric variation. We chose to use two different measures of pharyngeal jaw size, because they capture different aspects of pharyngeal jaw morphology: centroid size captures change in overall size, whereas weight is also affected by changes in bone density. For both the analyses of pharyngeal jaw size and shape, we removed the variation between time points before downstream analyses.

To analyse covariation between body and pharyngeal jaw shape, we used the Escoufier RV coefficient[38] and partial least squares analysis (PLS)[39] on the allometry-corrected shape variables. The Escoufier RV coefficient was used as a multivariate measure of association to test the null hypothesis of complete independence between body and pharyngeal jaw shape using the permutational procedure implemented in MorphoJ v1.06d[40]. If body and pharyngeal jaw shape were genetically independent, we would expect this statistical test to be nonsignificant in the F_2 mapping population. PLS was used to identify directions of maximal covariation between body and pharyngeal jaw shape. Only the first pair of axes—one for body, the other for pharyngeal jaw shape—was significant. We then used scores of each individual along each of these PLS axes for QTL mapping of body–pharyngeal jaw shape covariation.

A second set of individuals from each parental species, plus other three species of the Midas group, was used in the transcriptomic-based analysis to detect genes under selection. Although sequence evolution was analysed only in the two focal species, the other three species were included to generate a high-quality Midas reference transcriptome[41]. Two broods each from five Midas species were produced and sampled at 1 day post hatch (1 dph) and 1 month post hatch (1 mph): *A. astorquii* and *A. zaliosus* (crater Lake Apoyo), *A. amarillo* and *A. sagittae* (crater Lake Xiloá), and *A. citrinellus* (Lake Nicaragua). In total, nine samples per species were used for RNA extraction and RNA sequencing: three for the 1 dph stage and six for the 1 mph stage. Each of the 1 dph samples was obtained pooling three individual fish. The 1 mph samples were obtained using three fish and separating the head and the rest of the body (that is, head + body × 3 individuals = 6 samples). Individuals from the two different broods were included in the 1 dph and 1 mph samples.

For the population genomic analyses, we used a published data set[8] consisting of whole genome sequences of 26 pooled wild-caught individuals (PoolSeq) for each of *A. astorquii* and *A. zaliosus*.

Molecular methods and genotyping. Genomic DNA was extracted from the fin tissue of the two parentals, the two F_1 and 305 F_2 individuals using the Qiagen DNeasy Blood & Tissue Kit (Qiagen, Valencia, USA) following the manufacturer's protocol. The DNA quality of each sample was determined by agarose gel electrophoresis and quantified using a Qubit v2.0 fluorometer (Life Technologies, Darmstadt, Germany). Approximately 300 ng of DNA template of each sample was used to construct ddRAD[42] libraries following the modifications introduced in ref. 14. Seven ddRAD libraries, containing from 45 to 50 barcoded individuals each (see Supplementary Data 4 for details), were prepared and single-end sequenced in an Illumina HiSeq 2000 using four-colour DNA sequencing-by-synthesis technology with 101 cycles.

After barcode demultiplexing and filtering out low-quality reads with the 'process-radtags' script implemented in the Stacks v1.20 pipeline[43], a total of 8,748,063 (male) and 7,434,360 (female) sequences were obtained for the parents, 4,085,710 (male) and 3,799,752 (female) for the F_1, and an average of 2,396,189 sequences for the F_2 progeny (s.d. 984,832). Sequence length of each read was of 96 bp after removing its 5 bp barcode.

Genotyping was performed using the 'denovo_map.pl' module of Stacks with the parameters described in ref. 14, except for the coverage threshold to export SNPs in the Stacks' 'genotypes' script (-m) that was set at 10 and 15 in two different runs.

Linkage map construction. Linkage map construction was performed using the programme JoinMap v4.0 (ref. 44), which calculates genetic linkage maps in experimental populations of diploid species. To infer genetic linkage, we used both the methods implemented in JoinMap (the regression-based algorithm that uses the Kosambi mapping function and the Monte Carlo maximum likelihood mapping algorithm). This allowed us to identify inconsistencies between methods. A first linkage map was estimated to serve as a reliable backbone using a data set exported from Stacks that had higher coverage threshold (–m15). To increase marker density, the orders obtained with this data set were given as fixed orders to estimate the LGs with a genotype data set exported from Stacks that had a slightly less stringent coverage threshold of –m10 and pre-mapping filtering of loci with extreme segregation distortion (SD) levels ($P<0.005$). High levels of missing observations and SD can disturb the grouping phase of linkage map estimation[45].

To estimate a reliable backbone for adding more markers, markers with >20% missing data (> 61.6 missing genotypic observations) and under high levels of SD (χ^2: $P<0.010$) were excluded before grouping markers into LGs. Grouping was carried out using an independence LOD (logarithm of the odds, to the base 10) threshold of >5. The order of the markers in each LG was estimated using the maximum likelihood mapping function in JoinMap. Genotypes were visually inspected for all LGs and anomalous loci (those with a high incidence of double recombinations) were excluded.

The –m10 data set had a total of 1,766 loci, which was reduced to 512 by eliminating markers with >35% missing data or extreme levels of SD that indicate genotyping errors ($P>0.005$)[46]. The final map was produced by (i) assigning the markers ($n=410/512$) with up to 20% missing data and $P>0.01$ to LGs using a independence LOD cutoff of 5 (as above); (ii) assigning markers with a higher level of missing observations (between 20% and 35%) and SD ($0.01>P>0.005$) to the groups based on the strongest- cross-link values; (iii) giving the fixed orders estimated with the stringent data set; and (iv) excluding anomalous loci after inspecting the visual genotypes, congruence of the maximum likelihood and regression algorithms, the number of recombinations, incidence of improbable genotypes and nearest-neighbour stress values[47].

QTL mapping. We separately mapped body and pharyngeal jaw shape and PLS scores (covariation) using the multivariate version of Haley–Knott regression[48] on genotype probabilities implemented in the shapeQTL R package[49]. Body and pharyngeal jaw allometry-corrected shape variables were subjected to principal component analysis, to remove zero-variance dimensions before multivariate Haley–Knott regression. Using this multivariate approach, we could map all the variation in shape, as opposed to the projection on the between-group principal component we used previously[14]. Genotype probabilities were computed at 1 cM steps. A genome-wide significance LOD-score threshold was obtained using 1,000 random permutations under the null hypothesis of no association between the trait of interest and the genotype probabilities. For each QTL deemed significant at the 5% probability level, we estimated in shapeQTL the Bayesian credible interval for its position and its effect size in sum of squared deviations from the mean. Finally, we obtained predictions of QTL effect in terms of shape change vectors and PLS change estimates, which then we visualized in MorphoJ. Using a method to estimate statistical power in Haley–Knott regression[50], our sample sizes ($n=305$ for body shape and $n=265$ for pharyngeal jaw shape) would result in an estimated statistical power of 0.98 (body shape) and 0.96 (pharyngeal jaw shape), to detect a QTL explaining 5% of heritable phenotypic variance, under a type I error rate of 0.05 and a marker distance of 2.65 cM (that is, the average marker spacing in our map).

Finally, we used a recently developed approach[23] to test for the overlap of QTL regions for different traits. Given two sets of genomic regions and knowledge of the size of the LG, this approach uses random permutations to generate an empirical distribution of the number of overlaps between the two sets of genomic regions. The observed number of overlaps is then compared with this empirical distribution to obtain a *P*-value.

Transcriptomic analysis. FastPrep-24 homogenizer (MP Biomedicals) tubes were used to process 30 µg of each sample (30 s at 4.0 M). Total RNA from each sample (see paragraph 'Data sets used' above for details) was isolated using a Qiagen RNeasy Mini Kit with 100% ethanol used in all wash steps. RNA quality and quantity was assessed using a Bioanalyzer 2100 and a Qubit 2.0 fluorometer, respectively. Five-hundred nanograms of high-quality RNA (RNA Integrity Number value >8) was used to construct a barcoded sequencing library for each of the nine samples per species using the Illumina TruSeq RNA sample preparation kit (Low-Throughput protocol) according to the manufacturer's instructions (Illumina, San Diego, USA). To increase the average library insert size, chemical fragmentation was performed at 94 °C for 1 min. Paired-end sequencing of clustered template DNA was performed in an Illumina HiSeq 2500 with 309 cycles (151 cycles for each paired-read and 7 cycles for the barcode sequences).

After sequencing and pooling the different samples for each of the five species, we obtained 490,293,234 raw reads (from 81,089,742 to 115,093,936 reads per

species). Remaining adapters were removed and low-quality reads were filtered out using the software Trimmomatic v.0.32 (ref. 51) with default parameters and discarding sequences shorter than 50 bp. Reference-guided and *de novo* assembly approaches were performed using the filtered reads of the 45 samples combined. For reference guided assembly, we used the programme Stringtie v1.0.4 (ref. 52), setting the Midas genome as reference. TopHat v2.0.14 (ref. 53) and Bowtie2.2.3 (ref. 54) were used to map reads onto the Midas genome using default parameters. Samtools v1.2.1 (ref. 55) was used to convert the Bowtie output alignment from SAM to BAM format, to obtain the input file for Stringtie. The *gffread* utility implemented in the Cufflink v2.2.1 package[56] was used to extract the transcripts from the Midas genome. For the *de novo* assembly, the software Trinity v2.06 (ref. 57) was used to assemble the Midas reads (PasaFly transcript reconstruction mode, k-mer size of 32 and a minimum contig length of 200 bp). The obtained *de novo* and reference-based assemblies were combined and subjected to similarity searches (BLASTx v2.2.26 (ref. 58)) against the Nile tilapia (*Oreochromis niloticus*) protein data set (Ensembl release 73) using *e*-value $= 1e^{-10}$ as cutoff. The longest Midas transcript among those matching a unique tilapia protein was selected and its coding region was extracted according to the BLAST hit coordinates using bedtools v2.25.0 (ref. 59).

To infer orthology between the transcripts of the two focal species (*A. astorquii* and *A. zaliosus*), we implemented the following workflow. First, the extracted Midas cichlid coding regions were used as reference to independently align reads from the two focal species with CLC Genomics Workbench v6.5.1 (CLC bio, Aarhus, Denmark) with default parameters. Second, the consensus sequence was extracted from each alignment by exporting heterozygous sites, sites with low sequencing depth (threshold $10\times$) and low quality sites as unknown nucleotides (Ns). Finally, orthologue sequences were aligned with ClustalW v2.1 (ref. 60) and a custom Bash script was used to filter out sequence pairs with less than 30 complete codons. By explicitly coding heterozygous nucleotide sites within species as Ns and by removing shorter sequences, this workflow reduces the chances of finding false-positive genes under selection due to sequencing errors and decreases the probability of assuming that SNPs are fixed between species when they may actually be polymorphic within them.

This data set of 11,103 orthologue sequence pairs (where the alignable coding sequence length of the pairs ranged from 90 to 11,037 bp) was analysed to detect signatures of positive selection using the dN/dS approach[61] (the ratio between non-synonymous, dN, and synonymous, dS, substitutions). As estimation of dN and dS rates can be influenced by alignment methods, we used two different alignment methods (ClustalW[60] and the Needleman–Wunsch algorithm[62]), and three different approaches for the estimation of dN and dS. In particular, orthologous sequences aligned with ClustalW were analysed using separately the methods of Nei and Gojobori[63], and the Yang and Nielsen[64] with the PAML package v4.7a (ref. 65). The sequences aligned using the Needleman–Wunsch algorithm[62] with no penalty for sliding were analysed using the the method of Goldman and Yang[66], as implemented in the Matlab Bioinformatics Toolbox (Mathworks, Inc.). Out of the 11,103 orthologue sequence pairs, 298 had at least one polymorphic site (synonymous or non-synonymous; notice that codons containing ambiguous bases were excluded from the analysis). To avoid comparing paralogous genes, sequence pairs for which dS >0.1 were filtered out. This reduced the data set to 291 sequences. Exploratory scatterplots did not reveal any correlation between analysed sequence length (in codons) and dN, dS or dN/dS ratio. We considered and retained only sequences where dN/dS was consistently >1 in all three methods described above as candidate genes for selection. Finally, to further reduce the chances of false positives, all orthologue sequence pairs with dN/dS >1 were manually inspected to identify potential alignment errors. In all, this analysis suggested the presence of 71 candidate genes for selection between the two focal species.

Blast2GO[67] was used to perform the functional annotation of these 71 candidate genes. The same gene set was tested for overrepresentation of Gene Ontology terms by an enrichment analysis based on the Fisher's exact test (false discovery rate $= 0.05$) as implemented in Blast2Go. To test for Gene Ontology over-representation of the candidate genes relative to background, we compared them with the 11,103 genes that mapped uniquely to the tilapia proteins

Co-location of genes with dN/dS > 1 and QTL. The 49 RAD markers included in the QTL credibility intervals and the 71 candidate genes were aligned to the Midas draft genome v5 (ENA accession number ID PRJEB6974) using the BLASTn algorithm. For each query, the top BLAST hit was recorded and BLAST output parsed using a MySQL database. Given the length of the draft Midas genome scaffolds (maximum 8.1 Mb) and the average Bayesian confidence interval for the detected QTL (16.9 cM) and considering a genome size of 840 Mb, genes were considered as belonging to the QTL regions when both BLAST queries (RAD marker and gene under selection) were aligned to the same reference Midas scaffold.

Population genomics. We used low-coverage whole genome sequences for two pools of 26 wild-caught individuals for each of the two focal species[8]. A total of 69,161,026 (*A. astorquii*) and 92,024,490 (*A. zaliosus*) raw 151 bp paired-end reads were quality controlled using Trimmomatic. The filtered reads were then aligned to the Midas genome v5 using Bowtie and the output mapping files (BAM format) were then processed with Picard-tools v1.119 (http://picard.sourceforge.net.), to remove duplicates. Low-quality alignments (reads with mapping quality lower than 20, unmapped reads, reads in which both mates failed to align to the reference genome) were filtered out using the SAMtools *view* module. SAMtools *mpileup* module was used to extract SNP and coverage information from each pool. The average genome-wide coverage for the two species pools combined before and after the filtering steps applied was of $\times 15.2$ and $\times 9.2$, respectively (the distribution is shown in Supplementary Fig. 4). The PoPoolation2 v1.20127 (ref. 68) pipeline was then used to compute the fixation index (F_{ST}) and to perform a test of difference in allele frequencies (Fisher's exact test). More specifically, before computing these statistics, genomic sites were subsampled to the target coverage of ten, to avoid bias across sites produced by a non-uniform coverage (before subsampling, 48.5% of the Midas genome had a per-base coverage of at least $10\times$—see Supplementary Fig. 4). The minimum minor allele count at each site was set to 4, to drop only very rare alleles and to avoid overestimating heterozygous positions. The above-mentioned statistics were computed on non-overlapping windows of 500 bp with a minimum covered fraction of 0.5 (that is, at least half of the window has a coverage of $10\times$ and is included in the analysis), because using a sliding window approach reduces stochastic errors[69]. With these settings, we obtained 5,469 polymorphic sites (SNPs) covered by at least 10 reads in each species pool, which rendered 627 windows for which F_{ST} values were calculated. The *P*-values obtained with the Fisher's test were corrected for multiple tests using the binomial sequential goodness of fit procedure[70].

Co-location of QTL and population genomic data. The genomic scaffolds containing the RAD markers in the linkage map identified with the procedure described above (Co-location of genes with dN/dS > 1 and QTL) were also used in the interpretation of the results of the population genomic analyses. We computed the average F_{ST} for genomic scaffolds containing RAD markers in the QTL regions and the average F_{ST} for the genomic scaffolds containing the remaining RAD markers in the linkage map. We used the same principle to identify whether any of the QTL regions displayed significantly different allele frequencies.

Data availability. The population genomic data and the corresponding reference genome are available in the European Nucleotide Archive (accession numbers PRJEB6990 and PRJEB6974, respectively). The remaining data that support the findings of this study are available from the corresponding author upon request.

References

1. Mayr, E. *Animal Species and Evolution* (1963).
2. Via, S. Sympatric speciation in animals: the ugly duckling grows up. *Trends Ecol. Evol.* **16,** 381–390 (2001).
3. Coyne, J. A. Sympatric speciation. *Curr. Biol.* **17,** R787–R788 (2007).
4. Bird, C. E., Fernandez-Silva, I., Skillings, D. J. & Toonen, R. J. Sympatric speciation in the post 'Modern Synthesis' era of evolutionary biology. *Evol. Biol.* **39,** 158–180 (2012).
5. Michel, A. P. *et al.* Widespread genomic divergence during sympatric speciation. *Proc. Natl Acad. Sci. USA* **107,** 9724–9729 (2010).
6. Savolainen, V. *et al.* Sympatric speciation in palms on an oceanic island. *Nature* **441,** 210–213 (2006).
7. Barluenga, M., Stölting, K. N., Salzburger, W., Muschick, M. & Meyer, A. Sympatric speciation in Nicaraguan crater lake cichlid fish. *Nature* **439,** 719–723 (2006).
8. Elmer, K. R. *et al.* Parallel evolution of Nicaraguan crater lake cichlid fishes via non-parallel routes. *Nat. Commun.* **5,** 5168 (2014).
9. Bolnick, D. I. & Fitzpatrick, B. M. Sympatric speciation: models and empirical evidence. *Annu. Rev. Ecol. Evol. Syst.* **38,** 459–487 (2007).
10. Feder, J. L. & Nosil, P. The efficacy of divergence hitchhiking in generating genomic islands during ecological speciation. *Evolution* **64,** 1729–1747 (2010).
11. Via, S. Divergence hitchhiking and the spread of genomic isolation during ecological speciation-with-gene-flow. *Phil. Trans. R. Soc. B Biol. Sci.* **367,** 451–460 (2012).
12. Via, S., Conte, G., Mason-Foley, C. & Mills, K. Localizing FST outliers on a QTL map reveals evidence for large genomic regions of reduced gene exchange during speciation-with-gene-flow. *Mol. Ecol.* **21,** 5546–5560 (2012).
13. Elmer, K. R., Kusche, H., Lehtonen, T. K. & Meyer, A. Local variation and parallel evolution: morphological and genetic diversity across a species complex of neotropical crater lake cichlid fishes. *Phil. Trans. R. Soc. B Biol. Sci.* **365,** 1763–1782 (2010).
14. Franchini, P. *et al.* Genomic architecture of ecologically divergent body shape in a pair of sympatric crater lake cichlid fishes. *Mol. Ecol.* **23,** 1828–1845 (2014).
15. Franchini, P., Fruciano, C., Frickey, T., Jones, J. C. & Meyer, A. The gut microbial community of midas cichlid fish in repeatedly evolved limnetic-benthic species pairs. *PLoS ONE* **9,** e95027 (2014).
16. Schluter, D. Adaptive radiation in sticklebacks: size, shape, and habitat use efficiency. *Ecology* **74,** 699–709 (1993).

17. Svanback, R. & Eklov, P. Morphology dependent foraging efficiency in perch: a trade-off for ecological specialization? *Oikos* **102,** 273–284 (2003).
18. Rouleau, S., Glemet, H. & Magnan, P. Effects of morphology on swimming performance in wild and laboratory crosses of brook trout ecotypes. *Funct. Ecol.* **24,** 310–321 (2010).
19. Meyer, A. Cost of morphological specialization: feeding performance of the two morphs in the trophically polymorphic cichlid fish, *Cichlasoma citrinellum*. *Oecologia* **80,** 431–436 (1989).
20. Gavrilets, S., Vose, A., Barluenga, M., Salzburger, W. & Meyer, A. Case studies and mathematical models of ecological speciation. 1. Cichlids in a crater lake. *Mol. Ecol.* **16,** 2893–2909 (2007).
21. Flaxman, S. M., Wacholder, A. C., Feder, J. L. & Nosil, P. Theoretical models of the influence of genomic architecture on the dynamics of speciation. *Mol. Ecol.* **23,** 4074–4088 (2014).
22. Nosil, P. *Ecological Speciation* (Oxford Univ. Press, 2012).
23. Gel, B. *et al.* regioneR: an R/Bioconductor package for the association analysis of genomic regions based on permutation tests. *Bioinformatics* **32,** 289–291 (2016).
24. Gunter, H. M. *et al.* Shaping development through mechanical strain: the transcriptional basis of diet-induced phenotypic plasticity in a cichlid fish. *Mol. Ecol.* **22,** 4516–4531 (2013).
25. Rebsamen, M. *et al.* SLC38A9 is a component of the lysosomal amino acid sensing machinery that controls mTORC1. *Nature* **519,** 477–481 (2015).
26. Seehausen, O. *et al.* Genomics and the origin of species. *Nat. Rev. Genet.* **15,** 176–192 (2014).
27. Barluenga, M. & Meyer, A. Phylogeography, colonization and population history of the Midas cichlid species complex (*Amphilophus* spp.) in the Nicaraguan crater lakes. *BMC Evol. Biol.* **10,** 326 (2010).
28. Roesti, M., Hendry, A. P., Salzburger, W. & Berner, D. Genome divergence during evolutionary diversification as revealed in replicate lake–stream stickleback population pairs. *Mol. Ecol.* **21,** 2852–2862 (2012).
29. Fruciano, C., Franchini, P., Raffini, F., Fan, S. & Meyer, A. Are sympatrically speciating Midas cichlid fish special? Patterns of morphological and genetic variation in the closely related species *Archocentrus centrarchus*. *Ecol. Evol.* **6,** 4102–4114 (2016).
30. Slate, J. From Beavis to beak color: a simulation study to examine how much QTL mapping can reveal about the genetic architecture of quantitative traits. *Evolution* **67,** 1251–1262 (2013).
31. Boulding, E. G. *et al.* Conservation genomics of Atlantic salmon: SNPs associated with QTLs for adaptive traits in parr from four trans-Atlantic backcrosses. *Heredity (Edinb)* **101,** 381–391 (2008).
32. Boulton, K. *et al.* QTL affecting morphometric traits and stress response in the gilthead seabream (*Sparus aurata*). *Aquaculture* **319,** 58–66 (2011).
33. Bookstein, F. L. Landmark methods for forms without landmarks: morphometrics of group differences in outline shape. *Med. Image Anal.* **1,** 225–243 (1997).
34. Valentin, A. E., Penin, X., Chanut, J. P., Sévigny, J. M. & Rohlf, F. J. Arching effect on fish body shape in geometric morphometric studies. *J. Fish Biol.* **73,** 623–638 (2008).
35. Fruciano, C., Tigano, C. & Ferrito, V. Geographical and morphological variation within and between colour phases in *Coris julis* (L. 1758), a protogynous marine fish. *Biol. J. Linn. Soc.* **104,** 148–162 (2011).
36. Fruciano, C. Measurement error in geometric morphometrics. *Dev. Genes Evol.* **226,** 139–158 (2016).
37. Fruciano, C., Tigano, C. & Ferrito, V. Traditional and geometric morphometrics detect morphological variation of lower pharyngeal jaw in *Coris julis* (Teleostei, Labridae). *Ital. J. Zool.* **78,** 320–327 (2011).
38. Escoufier, Y. Le traitement des variables vectorielles. *Biometrics* **29,** 751–760 (1973).
39. Rohlf, F. J. & Corti, M. Use of two-block partial least-squares to study covariation in shape. *Syst Biol* **49,** 740–753 (2000).
40. Klingenberg, C. P. MorphoJ: an integrated software package for geometric morphometrics. *Mol. Ecol. Res.* **11,** 353–357 (2011).
41. Franchini, P., Xiong, P., Fruciano, C. & Meyer, A. The role of microRNAs in the repeated parallel diversification of lineages of Midas cichlid fish from Nicaragua. *Genome Biol. Evol.* **8,** 1543–1555 (2016).
42. Peterson, B. K., Weber, J. N., Kay, E. H., Fisher, H. S. & Hoekstra, H. E. Double digest RADseq: an inexpensive method for *de novo* SNP discovery and genotyping in model and non-model species. *PLoS ONE* **7,** e37135 (2012).
43. Catchen, J. M., Amores, A., Hohenlohe, P., Cresko, W. & Postlethwait, J. H. Stacks: building and genotyping loci *de novo* from short-read sequences. *G3: Genes, Genomes, Genetics* **1,** 171–182 (2011).
44. Van Ooijen, J. JoinMap 4. Software for the Calculation of Genetic Linkage Maps in Experimental Populations Kyazma BV, Wageningen, Netherlands (2006).
45. Haas, B. J. *et al.* *De novo* transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat. Protoc.* **8,** 1494–1512 (2013).
46. Henning, F., Lee, H. J., Franchini, P. & Meyer, A. Genetic mapping of horizontal stripes in Lake Victoria cichlid fishes: benefits and pitfalls of using RAD markers for dense linkage mapping. *Mol. Ecol.* **23,** 5224–5240 (2014).
47. Van Ooijen, J. W. & Jansen, J. *Genetic Mapping in Experimental Populations* (Cambridge Univ. Press, 2013).
48. Haley, C. S. & Knott, S. A. A simple regression method for mapping quantitative trait loci in line crosses using flanking markers. *Heredity* **69,** 315–324 (1992).
49. Navarro, N. R/shapeQTL: multiple QTL mapping for geometric morphometrics. *Université de Bourgogne, Dijon* (2014).
50. Hu, Z. & Xu, S. A simple method for calculating the statistical power for detecting a QTL located in a marker interval. *Heredity* **101,** 48–52 (2008).
51. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30,** 2114–2120 (2014).
52. Pertea, M. *et al.* StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.* **33,** 290–295 (2015).
53. Trapnell, C., Pachter, L. & Salzberg, S. L. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25,** 1105–1111 (2009).
54. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9,** 357–359 (2012).
55. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25,** 2078–2079 (2009).
56. Trapnell, C. *et al.* Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* **28,** 511–U174 (2010).
57. Grabherr, M. G. *et al.* Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* **29,** 644–U130 (2011).
58. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic Local Alignment Search Tool. *J. Mol. Biol.* **215,** 403–410 (1990).
59. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26,** 841–842 (2010).
60. Larkin, M. A. *et al.* Clustal W and Clustal X version 2.0. *Bioinformatics* **23,** 2947–2948 (2007).
61. Kimura, M. Preponderance of synonymous changes as evidence for the neutral theory of molecular evolution. *Nature* **267,** 275–276 (1977).
62. Needleman, S. B. & Wunsch, C. D. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* **48,** 443–453 (1970).
63. Nei, M. & Gojobori, T. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol. Biol. Evol.* **3,** 418–426 (1986).
64. Yang, Z. & Nielsen, R. Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. *Mol. Biol. Evol.* **17,** 32–43 (2000).
65. Yang, Z. PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24,** 1586–1591 (2007).
66. Goldman, N. & Yang, Z. A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol. Biol. Evol.* **11,** 725–736 (1994).
67. Conesa, A. *et al.* Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* **21,** 3674–3676 (2005).
68. Kofler, R., Pandey, R. V. & Schlotterer, C. PoPoolation2: identifying differentiation between populations using sequencing of pooled DNA samples (Pool-Seq). *Bioinformatics* **27,** 3435–3436 (2011).
69. Kofler, R. *et al.* PoPoolation: a toolbox for population genetic analysis of next generation sequencing data from pooled individuals. *PLoS ONE* **6,** e15925 (2011).
70. Castro-Conde, I. & de Una-Alvarez, J. Adjusted p-values for SGoF multiple test procedure. *Biom. J.* **57,** 108–122 (2015).

Acknowledgements

We thank C. Chang-Rudolf and L. da Conceicao Ferrao Beck for technical help, and M.L. Spreitzer for help with data collection. This work was supported by the ERC Advanced grant GenAdap 293700 by the European Research Council to A.M. C.F. was supported by a Marie Curie IEF Fellowship (grant number 327875—PlasticitySpeciation). P.F. was supported by the Deutsche Forschungsgemeinschaft (grant number FR 3399/1-1). V.K. was supported by the Czech Science Foundation (grant number 13-06264S).

Author contributions

C.F. carried out fish care, performed the morphometric analyses, QTL mapping, conducted molecular data analyses and wrote the manuscript. P.F. performed the laboratory work, conducted the molecular data analyses, QTL mapping and wrote the manuscript. V.K. performed molecular data analysis. F.H. built the linkage map. K.R.E.

carried out fish breeding and wrote the manuscript. A.M. conceived the study, caught and crossed the fish, and wrote the manuscript. C.F. and P.F. contributed equally to the study.

Additional information

Supplementary Information accompanies this paper at http://www.nature.com/naturecommunications

Competing financial interests: The authors declare no competing financial interests.

Reprints and permission information is available online at http://npg.nature.com/reprintsandpermissions/

How to cite this article: Fruciano, C. *et al.* Genetic linkage of distinct adaptive traits in sympatrically speciating crater lake cichlid fish. *Nat. Commun.* 7:12736 doi: 10.1038/ncomms12736 (2016).