MOLECULAR ECOLOGY

Molecular Ecology (2014) **23**, 5224–5240 doi: 10.1111/mec.12860

Genetic mapping of horizontal stripes in Lake Victoria cichlid fishes: benefits and pitfalls of using RAD markers for dense linkage mapping

FREDERICO HENNING,*‡ HYUK JE LEE,*†‡ PAOLO FRANCHINI* and AXEL MEYER*

Lehrstuhl für Zoologie und Evolutionsbiologie, Department of Biology, University of Konstanz, Universitätsstraße 10, Konstanz 78457, Germany

Abstract

The genetic dissection of naturally occurring phenotypes sheds light on many fundamental and longstanding questions in speciation and adaptation and is a central research topic in evolutionary biology. Until recently, forward-genetic approaches were virtually impossible to apply to nonmodel organisms, but the development of next-generation sequencing techniques eases this difficulty. Here, we use the ddRAD-seq method to map a colour trait with a known adaptive function in cichlid fishes, well-known textbook examples for rapid rates of speciation and astonishing phenotypic diversification. A suite of phenotypic key innovations is related to speciation and adaptation in cichlids, among which body coloration features prominently. The focal trait of this study, horizontal stripes, evolved in parallel in several cichlid radiations and is associated with piscivorous foraging behaviour. We conducted interspecific crosses between *Haplochromis sauvagei* and *H. nyererei* and constructed a linkage map with 867 SNP markers distributed on 22 linkage groups and total size of 1130.63 cM. Lateral stripes are inherited as a Mendelian trait and map to a single genomic interval that harbours a paralog of a gene with known function in stripe patterning. Dorsolateral and mid-lateral stripes were always coinherited and are thus under the same genetic control. Additionally, we directly quantify the genotyping error rates in RAD markers and offer guidelines for identifying and dealing with errors. Uncritical marker selection was found to severely impact linkage map construction. Fortunately, by applying appropriate quality control steps, a genotyping accuracy of >99.9% can be reached, thus allowing for efficient linkage mapping of evolutionarily relevant traits.

Keywords: adaptation genetics, colour-genes, ddRAD-seq, F-box, haplochromines, linkage mapping

Received 28 April 2014; revision received 2 July 2014; accepted 12 July 2014

Introduction

Deciphering the interplay between genotype and phenotype of an organism and its underlying mechanisms is one of the main goals in evolutionary biology (Streelman & Kocher 2000; Stinchcombe & Hoekstra 2008; Rockman 2012). Forward genetics is a powerful approach to elucidate the genetic underpinnings of ecologically or evolutionarily 'meaningful' phenotypes that are believed to be shaped by evolution through natural or sexual selection. But, so far, this technique has only been applied successfully to a handful of ecological model systems (Hoekstra *et al.* 2006; Linnen *et al.* 2009; Chan *et al.* 2010; Reed *et al.* 2011; Jones *et al.* 2012; Johnston *et al.* 2013).

In the post-genomic era through recent advances in genome sequencing technology, it has become possible for biologists to study the genetic basis of phenotypes

Correspondence: Axel Meyer, Fax: +49 7531 883018;
E-mail: axel.meyer@uni-konstanz.de
†Present address: Department of Biological Science, College of Science and Engineering, Sangji University, Wonju 220-702, Korea
‡These authors contributed equally to this work.

in ecologically or evolutionarily interesting species (Nadeau & Jiggins 2010; Henning & Meyer 2014). By performing controlled crossing experiments and mapping the genomic regions that harbour causal variants, it is now possible to begin to ask pertinent questions such as: Does the genetic architecture of a trait constrain its evolution (Schluter 1996)? Are coding or noncoding genomic regions more important in phenotypic evolution? Have similar phenotypes evolved through parallel evolution at the molecular level (Manceau *et al.* 2010; Elmer & Meyer 2011)? Are new mutations or standing genetic variation responsible for repeated evolution of similar traits (Domingues *et al.* 2012; Jones *et al.* 2012)? Does adaptive introgression play a role in young adaptive radiations (Loh *et al.* 2013)? Does adaptive phenotypic evolution typically result from a few genes of large effect or many genes with small effects (Orr 2005)? To most of these questions, general answers have not been reached yet, but through next-generation DNA technologies, the evolutionary community is poised to get there soon.

Determining the genomic architectures of adaptive phenotypes – the identification of the number, size and distribution of genomic locations responsible for the evolution of adaptive phenotypes – is a fundamental step towards the understanding of the basis of adaptive divergence (Bernatchez *et al.* 2010). More genomewide studies are required to link genotype and phenotype and particularly to advance our understanding of the genetic mechanisms leading to adaptive evolution. Once more of these kinds of studies have been conducted, it might be hoped that generally applicable answers will be reached.

The spectacular species richness of East African cichlid fishes and their striking diversity in morphology, coloration and behaviour have made them an ideal model for the study of adaptive evolution and speciation (Meyer *et al.* 1993; Kocher 2004; Henning & Meyer 2014). More than 1000 species can be found in the three large lakes of East Africa, Lakes Victoria, Malawi and Tanganyika. These cichlid adaptive radiations required only exceptionally short evolutionary time spans for their origin of these astonishingly diverse species flocks. In the case of the youngest Lake Victoria, more than 500 species evolved during <100 000 years (Verheyen *et al.* 2003; Elmer *et al.* 2009). Despite their diverse phenotypes, most East African cichlid species are still genetically extremely similar and therefore they often interbreed and can be hybridized in the laboratory. Until now, still very little is known about the genetic underpinnings of their speciation patterns and adaptive evolution, but studies have started to accumulate that are beginning to shed light on this issue (Kocher *et al.* 1998; Albertson *et al.* 2003; Streelman *et al.* 2003; Lee *et al.* 2005; Cnaani *et al.* 2008; Sanetra *et al.* 2009; Carleton *et al.* 2010; O'Quin *et al.* 2012, 2013; Recknagel *et al.* 2013; Henning & Meyer 2014). Many of the focal traits in these previous studies were found to be determined by a small number of genes, but it is unclear whether this result reflects the recent recruitment of major loci or experimental artefacts (Rockman 2012; Slate 2013).

Horizontal stripes in cichlid adaptive radiations

Undoubtedly, body coloration is one of the traits that most strongly affects cichlid adaptation and speciation (Kocher 2004). Sexual selection by female mating preference for male colour polymorphisms (i.e. nuptial coloration) is believed to play a large role in the exceptionally rapid genesis of new species in adaptive radiations of the East African cichlids (Seehausen *et al.* 1999; Seehausen 2000). Marked sexual dimorphisms and exuberant male colour patterns in species-rich lineages provide evidence that sexual selection might be a strong force driving the evolutionary diversification of the cichlid fishes (Maan *et al.* 2004).

The evolution of stripe patterns in the East African cichlids has been suggested to be driven by ecological adaptations to their habitat environments (Seehausen *et al.* 1999). The vertical bar patterns are mostly found in cichlids inhabiting structurally complex habitats such as rocky substrates and vegetation, which suggests the evolution of background matching as a means of predator avoidance. Horizontal stripes have evolved repeatedly and in parallel in all major cichlid fish radiations (Fig. 1A); however, this phenotype is restricted to relatively few genera and species by comparison with vertical bars that are found in the majority of cichlid lineages and species. The evolution of horizontal stripes is suggested to be associated with a piscivorous (i.e. feeding on fishes) feeding mode and/or shoaling behaviour (Seehausen *et al.* 1999). Shoaling behaviour is probably related to predation avoidance and predator intimidation (Baerends *et al.* 1986). Lateral stripes have been defined as mid-lateral and dorsolateral (Fig. 1B). These two stripes co-occur in most species that have lateral stripes (Seehausen *et al.* 1999).

Dense linkage map construction in nonmodel organisms

The development of reduced-representation library techniques, in particular ddRAD-seq (double-digest restriction-associated DNA sequencing), has been received with excitement by the research community interested in genetic mapping on nonmodel organisms (Davey *et al.* 2011). These methods allow for a cost- and time-efficient construction of dense linkage maps with

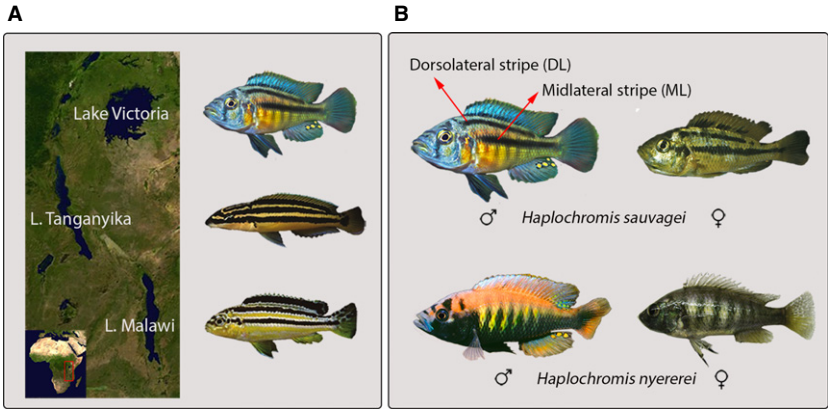


Fig. 1 Horizontal stripes in cichlid radiations. (A) Horizontal stripes are present in all major African cichlid radiations. Representatives shown are males of *Haplochromis sauvagei* (Lake Victoria), *Julidochromis ornatus* (Lake Tanganyika) and *Melanochromis auratus* (Lake Malawi). (B) Phenotypes mapped and species used in the experimental cross of the present study. Pictures were kindly provided by Erwin Schraml (*H. sauvagei*) and Ad Konings (all others).

no prior genomic information. These markers can be readily used for linkage map construction (e.g. Amores *et al.* 2011; Recknagel *et al.* 2013), genetic mapping of both mendelian and quantitative traits in line cross experiments (e.g. Franchini *et al.* 2013) and also to scaffold draft genome sequences (O'Quin *et al.* 2013) or even nearly complete genomes (Roesti *et al.* 2013). The approach is based on the simultaneous discovery and genotyping of tens of thousands of small sequence tags selected by flanking restriction sites (Miller *et al.* 2007; Baird *et al.* 2008) followed by the filtering for informative and mappable loci using bioinformatic pipelines such as Stacks (Catchen *et al.* 2011). Further developments of the RAD-seq method involve the use of two restriction enzymes to increase the efficiency and the correlation of sampled genomic regions across individuals (Peterson *et al.* 2012)3.

The high incidence of missing genotype data and systematic errors in RAD-seq data calls for increased attention in quality control steps (Ward *et al.* 2013). Genotyping errors can occur systematically and have a severe impact on linkage map estimation leading to spurious linkages, incorrect ordering and estimation of interval sizes (van Ooijen & Jansen 2013). Markers inflicted with systematic errors can be identified because they deviate strongly from the expected Mendelian segregation proportions (i.e. segregation distortion, SD). Many investigators might be reluctant to discard loci under SD because this can also have interesting biological causes (Lyttle 1991; Phadnis & Orr 2009; Kozielska *et al.* 2010). However, extreme values of SD are usually an indication of methodological artefacts. It is important to note that the problems caused by both missing data and incorrect genotypes are aggravated with increasing marker density (Feakes *et al.* 1999; Jansen *et al.* 2001; Hackett & Broadfoot 2003). Therefore, investigators working with high-throughput SNP data should be particularly aware of these issues because linkage maps constructed with them are characteristically dense and can have a high incidence of missing data (Ward *et al.* 2013). Genotyping errors are rarely quantified, and strategies for identifying and dealing with them have only seldom been addressed in the literature (Bonin *et al.* 2004; Pompanon *et al.* 2005).

Corrections made after online publication August 19, 2014: Fig.1 label has been updated.

Aims and hypothesis

The aims of this study were twofold. The primary aim was to understand the genetic segregation and to map lateral stripes, an ecologically relevant trait in cichlid radiations. The evolutionary significance of body stripe patterns in cichlid phenotypic adaptations has long been recognized, and only little attention has been paid so far to the genetic basis underlying those traits. In addition, we addressed a more general methodological issue regarding the scoring and mapping of RAD tags. Our experimental design was used to empirically estimate the error rate in a RAD data set and assess the issues to be cautious of when estimating dense linkage maps.

Materials and methods

Experimental cross set-up and phenotyping

The following experimental crosses between closely related species of the East African haplochromine cichlid fishes from Lake Victoria, which differ markedly in their body stripe patterns, were attained for this study: a male *Haplochromis nyererei* and a female *H. sauvagei* (Fig. 1). *H. sauvagei* have two horizontal stripes including the dorsolateral stripe (DL) and the mid-lateral stripe (ML) that are absent in *H. nyererei*. These two species were caught in the wild and subsequently reared separately for several generations under laboratory conditions. A single male *H. nyererei* and five females of *H. sauvagei* were initially kept in a 200-L

aquarium to generate F_1 hybrids. When a female was mouth-brooding, she was carefully transferred to a smaller 120-L tank. The P female (parental) was isolated into a different tank once juvenile fish were released from the mouth. The resulting F_1 generation was raised to sexual maturity (approximately 6–9 months), and groups of several females each with a single male were established for the F_1 intercross. Young fry of the F_2 generation, comprising 9–43 individuals (mean family size ≈22), was further obtained as done for the F_1 fish. All the F_2 hybrids were raised in the same size aquaria (200 L) under identical conditions to minimize environmental effects on their phenotypes. A total of 196 F_2 offspring resulted from intercrossing one male with nine different females. However, the final mapping population included only 171 F_2 individuals along with 10 F_1 and two P individuals.

The presence/absence of horizontal stripes (DL and ML) was scored when the F2 individuals were adult and sexually mature at about 9 months of age. All the fish were photographed in a standardized manner, and a fin clip was taken and stored in 98% ethanol for genetic analyses.

DNA extraction and library preparation

Genomic DNA was isolated from a small piece of fin tissue for the two P, 10 F_1 and 171 F_2 individuals using Qiagen DNeasy Blood & Tissue Kit (Qiagen, USA) following the manufacturer's recommendations including the RNase treatment to eliminate residual RNA. The DNA integrity of every sample was evaluated by agarose gel electrophoresis and quantified using a QUBIT v2.0 fluorometer (Life Technologies, Germany).

Approximately 300 ng of DNA template of each sample was double-digested using the restriction enzymes PstI-HF and MspI (New England BioLabs, Beverly, MA) in a single reaction. The double-digest restriction site-associated DNA (ddRAD) library preparation protocol followed the methodology described in Peterson *et al.* (2012). Briefly, after digestion, each sample was ligated to barcoded Illumina primers and then pooled at equimolar concentrations. A total of five RAD libraries were prepared, including 34, 44, 44, 43 and 43 individuals, respectively. Each F2 individual was included in only one of the five libraries (171 F2s). The parents were included in five libraries (Table S1, Supporting information). The libraries were size-selected for a range of 335–405 bp using a Pippin Prep® electrophoresis system (Sage Science, Beverly, USA). The ddRAD libraries were run on an Illumina HiSeq2000 at the Tufts University Genomics Center in Boston (TUCF Genomics) using a single-end protocol with 101 cycles.

Marker discovery, genotyping and linkage mapping

Candidate RAD loci were identified in the Illumina raw reads using the open-source STACKS PIPELINE v0.99999 (Catchen *et al.* 2011). Sequences of each individual were grouped by barcode, cleaned from low-quality reads and truncated to a length of 96 bp using the Stacks script 'process_radtags'. The program BOWTIE v1.0.0 (Langmead *et al.* 2009) was used to align the processed reads to the *Haplochromis nyererei* draft genome (Brawand *et al.* 2014) allowing two mismatches. Only reads that aligned uniquely to the reference genome (exported in 'sam' format) were used as input in the 'ref_map.pl' Stacks Perl script. This script executes three programs in succession that build loci in each sample, based on the reference alignment, and call SNPs in each (pstacks), create a catalogue of loci for the parents and the progeny (cstacks) and match each catalogue of loci of the sample against the parent catalogues (sstacks).

The ref_map.pl parameters were set as the default except for the following parameters: minimum depth of coverage to report a stack in pstacks (-m = 3) and number of mismatches allowed between loci when building the catalogue (-*n* = 3). Finally, the Stacks script 'genotypes' was used to export different progeny genotype data sets with different combinations of 'm' – minimum stack depth required before exporting a locus in a particular individual – and 'r' – minimum number of progeny required to print a marker.

We generated a series of genotype tables using the genotypes module of Stacks in which the -m (minimum coverage per individual) parameter was varied from 5 to 30. The final data set was chosen based on our estimation of genotyping accuracy and exported from Stacks using a minimum coverage threshold of 20× and had 1517 markers. A set of 1517 loci that were exported using the Stacks genotype module were further analysed using JOINMAP 4 (Van Ooijen 2006).

Empirical estimation of genotyping accuracy

To estimate the occurrence of genotyping errors, we took advantage of having used 10 different F1 individuals for establishing the F2 mapping panel. Loci where all the 10 F1s were heterozygotes are most likely homozygous in the parents (there was no case of both parents being heterozygotes, abxab). For loci in which the parents are homozygotes for different alleles (i.e. aaxbb), all F1s are expected to be heterozygotes (ab). Therefore, the frequencies of genotypes that are heterozygous in P provide an estimate of false heterozygotes and loci that were homozygous in F1 were used to infer the frequency of false homozygotes.

Premapping quality control and marker selection

Two rounds of quality control were performed based on the recommendations in van Ooijen & Jansen (2013). Premapping quality control was done before linkage map construction and consisted in excluding (i) individuals with > 30% missing data, (ii) markers exhibiting >20% missing data and (ii) markers under highly significant SD ($P < 0.001$).

Linkage mapping and post-mapping QC

Markers were grouped on linkage groups (LG) using an independence LOD threshold of 6. Using this threshold, 22 groups were obtained which matched the chromosome number of these species (Mazzuchelli *et al.* 2012). The final map was generated using the regression algorithm and Kosambi mapping function (as it is required by the trait mapping software). In the cases where the order obtained from maximum likelihood was more probable, it was given as a fixed order. The 22 groups were then ordered using the Kosambi regression mapping function screened for anomalous loci and incorrect orders using the strategy detailed below.

Agreement of estimation methods. As a first assessment of the quality of the linkage maps, LGs were estimated using two different algorithms, namely the regression (with Kosambi mapping function) and maximum likelihood. Comparing the distances and orders of the LGs estimated using both methods provides a first step to spot potential anomalous loci and problematic LGs (van Ooijen & Jansen 2013).

Inspection of genotype probabilities. The probability of double recombinants is increasingly small over shorter intervals and was used to identify errors in ordering and genotyping.

Interval and map size. The total size of the linkage groups were compared after exclusion of potentially erroneous markers. In general, shorter maps are an indication of higher quality, because errors are treated as recombination and their effect is to inflate intervals between markers. Missing data can also lead to incorrect estimation of distances. The maximum-likelihood algorithm implemented in JoinMap allows for the placement of adjacent markers at distances much higher than 50 cM. Naturally, these can only occur by error (i.e. flipping of paternal and maternal alleles) and this was used as a tool to identify potential erroneous markers.

Segregation distortion (SD). Loci under significant SD that were retained in premapping QC ($P > 0.001$) were further evaluated as follows. Each LG was compared before and after removing the markers under SD, and the markers that distorted the map were excluded from further analysis. Marker exclusion criteria involved (i) the inflation of map distances, (ii) incidence of improbable genotypes (e.g. double recombinants), (iii) drastic changes of orders and (iv) high levels of nearest neighbour stress.

Trait mapping

Mapping was done in two steps. First, a genomewide map was constructed and the genomic region associated with the inheritance of horizontal stripes was identified. We then determined the candidate interval (minimum region flanked by recombinants) and narrowed down the interval by recombinant breakpoint analysis using additional RAD markers that mapped within this region.

Linkage mapping was performed for the focal trait (presence of horizontal stripes) using an interval mapping (IM) algorithm implemented in the software Windows QTL CARTOGRAPHER V2.5_011 (Wang *et al.* 2012). This algorithm is an extension of the originally described IM method (optimized for continuously distributed traits) and is specifically designed to map QTL in binary and ordinal traits. Analysis was conducted with a 1.0 cM increment. LOD thresholds for testing the significance of QTL peaks were calculated using 1 000 permutations and a significance level of $P < 0.05$.

A less stringent marker selection was exported using the genotype module of Stacks (-m 5 and -r 70) to search for markers that map within the candidate interval. This strategy is based on the observation that systematic sources of missing data and incorrect genotypes are more likely to cause errors in the grouping stage of linkage map estimation. Fine mapping involves only focusing on the few recombinants identified for recombinant breakpoint analysis in a way that errors and missing data in other individuals will not affect this analysis (provided that the markers can be placed in the interval with confidence). When narrowing down on particular intervals in one or few LGs, it becomes possible to visually inspect the ordered genotypic data and decide whether to exclude or reorder markers.

Genomics and annotation of candidate interval

The cichlid genomes that were used to annotate the candidate interval are described in Brawand *et al.* (2014) and are available at http://cichlid.umd.edu/cichlidlabs/kocherlab/genomebrowsers.html. The Nile tilapia (*Oreochromis niloticus*) genome (Orenil1.0, Ensembl v72) is

currently the best annotated cichlid genome and was used as a reference to annotate the detected genomic interval. The Tilapia anchored genome assembly v.1.1 was used for comparative genomics because it is the only cichlid genome that is anchored to linkage groups. To retrieve the genome sequences, the RAD loci delimiting the candidate interval identified by the fine mapping analysis were aligned to tilapia genome using the BLASTn algorithm and the region spanning the BLAST hits was retained. The genes included in this region were retrieved using the Ensembl BioMart web-tool.

Results

Interspecific cross and phenotyping

The inheritance of stripes in F2 followed the expected proportions of a single-locus trait and shows that the presence of the horizontal stripe is recessive. Forty-one striped and 130 nonstriped F2s were obtained which is not significantly different from the expected Mendelian proportion of 1:3 ($\chi^2 = 0.236$, d.f. = 1, $P = 0.627$), and the horizontal stripes were completely absent in the F1s. Family-level analysis of F2s confirmed this pattern, that is, all the 9 F2 families followed a Mendelian ratio of 1:3 ($P > 0.05$). All of the striped F2s presented both the mid-lateral (ML) and the dorsolateral (DL) stripes, indicating that these two stripes segregate as a single Mendelian trait.

Sequencing output

After the cleaning pipeline was implemented, 454144120 sequences across the five RAD libraries were obtained (averaging 90828824 reads per library). For the female and male parental samples, 7235653 and 4902683 reads were obtained, respectively. For the 10 F_1 hybrids (one male and nine females), the average number of reads per individual was 2666946. The F_2 progeny libraries contained 2428867 reads per individual (Table S1, Supporting information).

Linkage map construction

From a total of 1517 loci exported from Stacks, 906 passed the premapping QC filtering and 867 were retained after post-mapping QC (Table S2, Supporting information). We obtained a genetic linkage map consisting of 867 SNP markers on 22 linkage groups spanning 1130.63 cM and mean marker spacing of 1.30 cM (Table 1, Fig. 2). The distribution is, however, skewed and the median spacing is 0.5 cM. Over 63% of the distances between adjacent markers are equal to or <1 cM (Fig. S1, Supporting information).

Table 1 Descriptive statistics of the *Haplochromis nyererei* and *H. sauvagei* hybrid genetic linkage map. Numbers in brackets show the proportion of markers that gave consistent mapping results when blasted against the Tilapia anchored genome

Linkage Group	Homologous LG in Tilapia	Size (cM)	Number of markers
1	7 (32/33)	52.406	61
2	6 (34/34)	55.088	56
3	1 (28/32)	49.146	56
4	12 (38/40)	52.761	52
5	17 (28/29)	53.673	55
6	3 (16/17)	49.59	41
7	16–21 (29/29)	56.633	44
8	11 (29/30)	52.278	41
9	2 (24/26)	55.543	43
10	23 (11/12)	56.073	37
11	5 (21/21)	57.836	39
12	22 (22/24)	47.354	39
13	15 (17/18)	52.14	38
14	20 (14/14)	36.081	28
15	19 (28/28)	55.481	38
16	14 (16/16)	53.056	34
17	18 (19/19)	45.896	31
18	4 (18/19)	50.987	29
19	10 (13/16)	48.781	28
20	8–24 (22/23)	46.3	29
21	9 (14/15)	51.265	29
22	13 (14/14)	52.262	19
Total		1130.63	867

A total of 71.6% ($n = 621$) of the RAD tags could be unambiguously mapped by blast (bit score > 100) to the Tilapia draft genome. From these 621 markers, 112 markers mapped to unplaced scaffolds and the remaining 82% ($n = 509$) mapped to the 22 tilapia linkage groups (Table 1). Only 4% of the markers mapped inconsistently to tilapia linkage groups. For instance, 32 of the 33 markers on LG1 with unambiguous blast results (present map) mapped to tilapia LG7, with the exception of one marker. These are likely to reflect cases where the RAD tags are repetitive. Mapping to paralogous regions will occur if the homologous regions either (i) have evolved faster than the paralogous ones or (ii) are missing from the current draft assembly.

Genetic mapping of horizontal stripes

A single region on linkage group 17 was highly associated with the inheritance of lateral stripes with LOD >37 (Fig. 3A). The analysis of the recombination breakpoints shows that the causal polymorphism is located between markers 90597 and 84385 (Fig. 3). This interval was refined by including four additional markers within the interval and was finally reduced

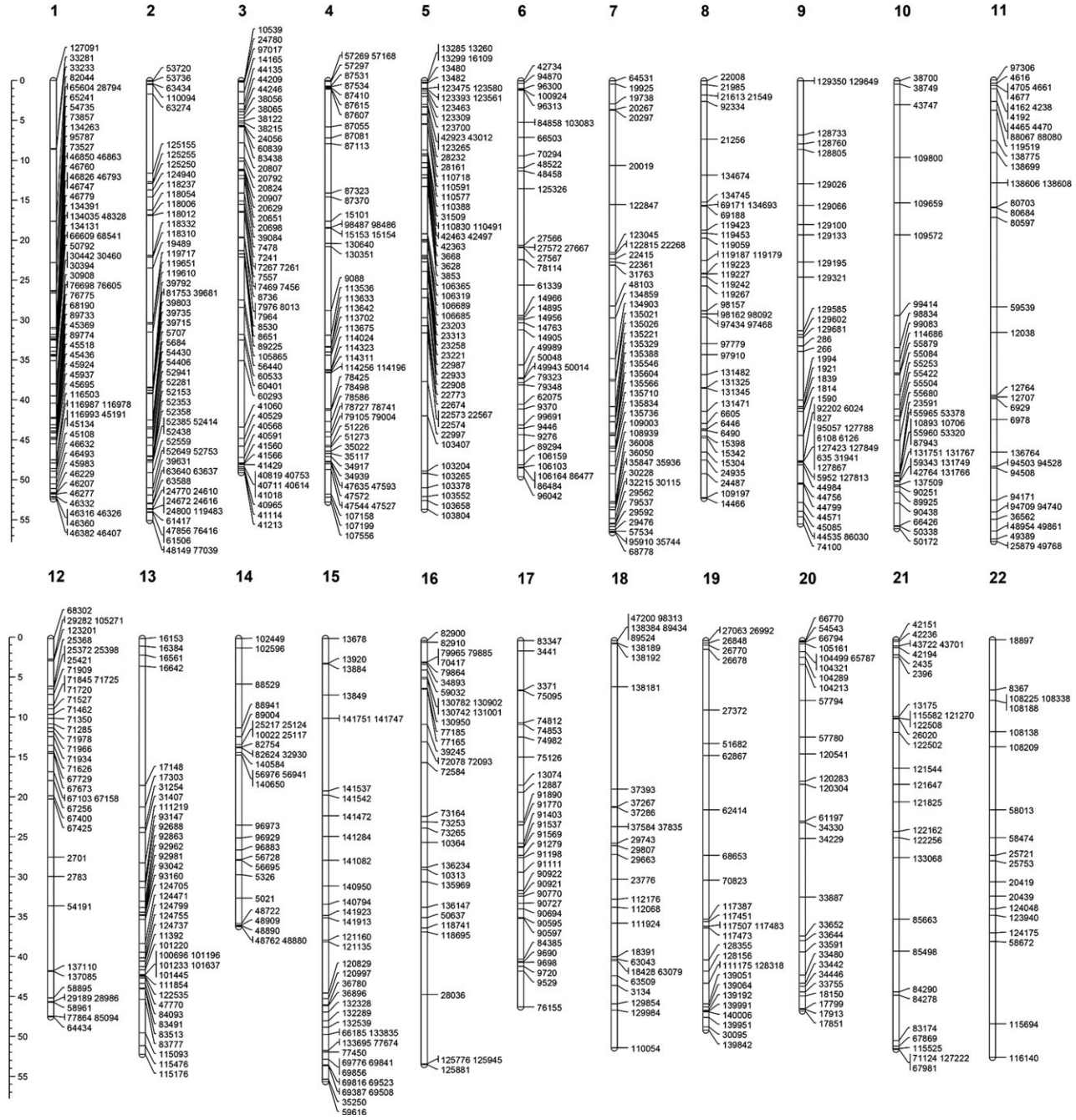


Fig. 2 Genetic linkage map of hybrids between *Haplochromis nyererei* and *H. sauvagei*. Distances in cM estimated using the Kosambi mapping function are given by the scales on the left.

to a region of <2 cM between markers 90805 and 84385 (Fig. 3B).

Genomics and annotation of candidate interval

The two RAD markers that flank the candidate interval map contiguous sequences in the *Metriaclima zebra* (scaffold 73), *Neolamprologus brichardi* (scaffold 10) and Tilapia genomes (LG18) where the distance between them is 559Kb, 580Kb and 598Kb, respectively. The region in Tilapia contains 35 annotated genes (Table 2). The size of the genomic interval between markers 90805 and 84385 could not be directly determined in the *H. nyererei* draft genome sequence because each marker mapped to two scaffolds: 90805 maps to positions 277439–277533 on scaffold 3 (8 995 252 bp of total length) and 84385 maps to positions 322473–322567 on scaffold 361 (447 939 bp total length). Scaffold 361 has over 25Kb of missing sequence

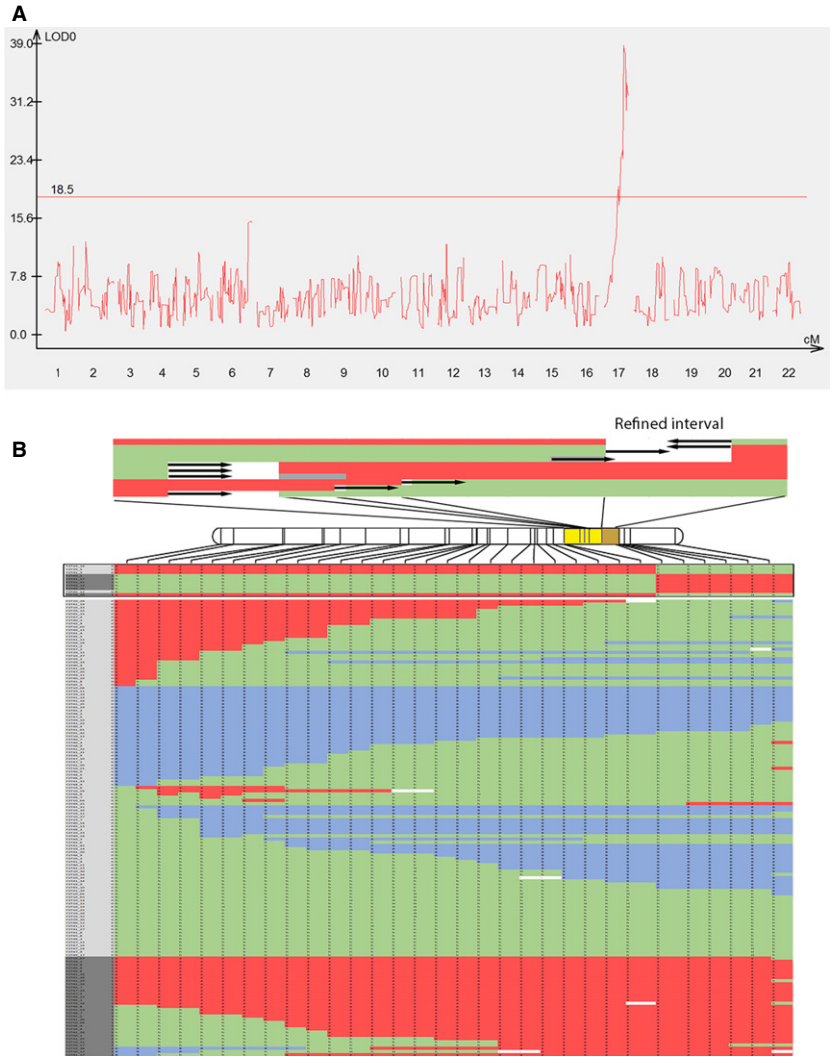


Fig. 3 Genetic mapping of lateral stripes. (A) Genomewide categorical trait mapping identifies a single region on LG17 with highly significant association with the inheritance of lateral stripes. (B) Recombination Breakpoint analysis and fine mapping. Each line represents the genotypes of one F2 individual across LG17. Nonstriped F2s are coded light grey and striped F2s are coded dark grey. Markers homozygous for the striped and nonstriped parents are coded in red and blue, respectively. Loci that are heterozygous are coded green. The 10 F2s outlined in black and located on the top are recombinants in the genomic region identified in the genomewide analysis (shown as gold and yellow in the schematic representation of LG17) and provided information for fine mapping the locus responsible for the inheritance of lateral stripes to the region shown in gold following the addition of four markers.

in its 5′ end that probably precluded joining with scaffold 3 in the genome assembly. A single gene within this interval, TBLX1, has a known function in stripe patterning (Kawakami *et al.* 2000).

Linkage mapping quality control

Coverage threshold. The most important source of random errors is the occurrence of false homozygotes due to sampling a single allele from a heterozygote (Table 3). The coverage threshold was the parameter that strongly affected the frequency of genotyping errors but did not affect the occurrence of systematic allelic dropout. The error rate was still above 2% even when using a high cut-off of 30× coverage.

The re-estimation of the proportion of incorrect genotypes in the 10 F1s was done with the 867 markers that were retained after pre- and post-mapping QC. We found that the number of false heterozygotes and false homozygotes dropped to zero and six (0.08%), respectively (Table S2, Supporting information). Four loci (20419, 23591, 124048, 129066) had instances of false homozygotes in the F1s. Curiously, two of these (20419 and 124048) had errors in two different F1s, thus indicating that these markers are more prone to errors. Excluding these yields a proportion of errors of only 0.026%, which is much closer to the expected occurrence of allelic dropout by chance.

Segregation distortion. Post-mapping quality control showed that the impact of poorly genotyped loci can indeed be drastic. Several linkage groups were severely distorted with the addition of markers under SD (Fig. 4).

Missing data. In many cases, genetic distances greater than zero were inferred even in the total absence of detected recombination events (Fig. S2, Supporting information). Although the inferred distances were small, they led to the inference of an order that is not supported by a single recombination event. We found that the

Table 2 Positional candidate genes. Ensembl annotations were retrieved for the homologous tilapia genomic scaffold of 598Kb between markers 90805 and 84385

Ensembl gene ID	Sequence description	Gene symbol
ENSONIG00000010618	arf-gap domain and fg repeat-containing protein 1-like	AGFG1
ENSONIG00000010622	f-box only protein 36	FBXO36
ENSONIG00000010623	g protein-coupled receptor kinase 7a-like	GRK7
ENSONIG00000010624	Sodium/potassium-transporting atpase subunit beta-3-like	ATP1B3
ENSONIG00000010628	Glycerol kinase 5	GK5
ENSONIG00000010631	Nucleolin-like isoform ×2	–
ENSONIG00000010633	Probable cation-transporting atpase 13a3-like	–
ENSONIG00000010641	Transcription factor hes-1	HES4
ENSONIG00000010642	Carboxypeptidase n subunit 2	CPN2
ENSONIG00000010643	Claudin-1	CLDN1
ENSONIG00000010655	b-cell lymphoma 6 protein isoform ×1	BCL6
ENSONIG00000010657	ump-cmp kinase	CMPK1
ENSONIG00000010659	Protein fam78b	FAM78B
ENSONIG00000010660	udp-n-acetylglucosamine transporter	–
ENSONIG00000010661	u3 small nucleolar ribonucleoprotein protein imp3	IMP3
ENSONIG00000010664	Vesicle-trafficking protein sec22b-like	–
ENSONIG00000010667	riken cdna 1700025 g04 gene	C1orf21
ENSONIG00000010668	krueppel-like factor 9-like	KLF–14
ENSONIG00000010669	PREDICTED: uncharacterized protein LOC101474939	–
ENSONIG00000010670	v-type proton atpase subunit d 1	ATP6V0D2
ENSONIG00000010672	Sorting nexin-16	SNX16
ENSONIG00000010673	Receptor-interacting serine/threonine-protein kinase 2	RIPK2
ENSONIG00000010677	y+l amino acid transporter 2-like	–
ENSONIG00000010679	Coiled-coil domain-containing protein 39	CCDC39
ENSONIG00000010683	dcn1-like protein 1	DCUN1D1
ENSONIG00000010685	Eukaryotic initiation factor 4a-ii	EIF4A2
ENSONIG00000021439	Lactosylceramide -n-acetyl-beta-d-glucosaminyltransferase	B3GNT5
ENSONIG00000010689	f-box-like wd repeat-containing protein tbl1xr1-like	TBL1XR1
ENSONIG00000010692	Calcium-activated potassium channel subunit beta-2-like	KCNMB1
ENSONIG00000010693	Phosphatidylinositol-bisphosphate 3-kinase	PIK3CA
ENSONIG00000010699	Calcium-activated potassium channel subunit beta-3-like	KCNMBS
ENSONIG00000010705	Ephrin type-b receptor 3-like	EPBHB3
ENSONIG00000010713	Neuromedin-u receptor 1-like	NMUR1
ENSONIG00000010715	Deoxyribodipyrimidine photo-lyase-like	–
ENSONIG00000010717	uap56-interacting factor	FYTTD1

impact of missing data and SD differs depending on the algorithm used for linkage mapping. The ML algorithm implemented in JoinMap was more robust leading to correct orderings and complete linkage of markers more often than the regression algorithm (Fig. S2, Supporting information). Furthermore, the presence of markers with more than 20% missing genotypes led to difficulties in the grouping stage of linkage map construction. Two LGs remained linked at very high LOD scores (>30) but broke down to separate LGs (that are supported by the comparison to tilapia) at a lower LOD score of 4 with the removal of the markers that had >20% missing data.

Discussion

Our results show that stripes in *H. sauvagei* are inherited recessively and that the ML and DL stripes are both governed by the same single locus. By developing a dense linkage map with high-quality RAD-seq markers, we were able to map both ML and DL to a single genomic interval on LG17. This genomic interval contains members of the F-box gene family that was shown to be under strong positive selection in the cichlid radiations (Terai *et al.* 2002) and that causes the disruption of stripe patterns when knocked-out in zebrafish (Kawakami *et al.* 2000).

We empirically quantified for the first time the error rates in RAD-seq data sets, and our results highlight that those using RAD-seq should be aware of the potential problems caused by genotyping errors and missing genotype observations. Quality control in both premapping and post-mapping phases are successful in achieving genotyping accuracy >99% while still retaining a number of genetic markers that exceeds the

resolution given by the number of progeny in most studies.

The genetic basis of horizontal stripes

The simultaneous presence of both mid-lateral and dorsolateral stripes in most cichlid species (Seehausen *et al.* 1999) is most likely due to both being determined by a single locus. The complete absence of F2 individuals with only either of the two stripes in our mapping panel supports this hypothesis. This would indicate an interesting genetic constraint in the evolution of stripe patterns. Although the genomic interval to which ML and DL were mapped to is still relatively large (>500Kb in tilapia) and contains numerous positional candidate genes (n = 35), two of them are promising candidates for being the causal gene, *F-box 36* (FBXO36) and *F-box-like WD repeat-containing protein* (TBLXR1). The WD-repeat domains have diverse regulatory functions that include mediating protein–protein interactions (Stirnimann *et al.* 2010) and are believed to be associated with the determination stripe patterns (Terai *et al.* 2002). *Hagoromo* (HAG), another F-box/WD40, was shown to disrupt stripe patterning in zebrafish when knocked-out (Kawakami *et al.* 2000). Interestingly, the surface of the WD-repeat domain has evolved quickly through positive natural selection in the haplochromine lineage (Terai *et al.* 2002). Further work on fine mapping this interval and functional validation is currently being pursued.

The Mendelian basis of the lateral stripe phenotype makes it an ideal phenotype for further work on determining the causal variants and for addressing specific hypothesis pertaining the evolution of this trait. Of course, the fact that this trait is Mendelian in *H. sauvagei* does not imply it is so in other species bearing similar phenotypes and further forward-genetic projects should be conducted to fully understand the evolution of this trait in the wild. Future studies should investigate whether the same genomic region is associated with determining lateral stripes in other species. If the causal variant can be identified, it might be possible to test whether adaptive introgression has played a role in cichlid adaptive radiation and whether this causal variant has been transported to the radiations by riverine species (Loh *et al.* 2013). One strategy for investigating these questions is *functional phylogenomic*s (Henning & Meyer, in press): contrasting gene trees from nonlinked regions of the genome with those from the causal region (e.g. Jones *et al.* 2012). Another approach is the forward-genetic mapping in additional species that bear stripe patterns. Genetic mapping using species that differ in the number or colour of lateral stripes could reveal whether additional genes or variants of the same gene were recruited in the evolution of this pattern. Some naturally occurring varieties or species lend themselves particularly well to these follow-up studies and would allow the test of the hypothesis we put forth in the present study.

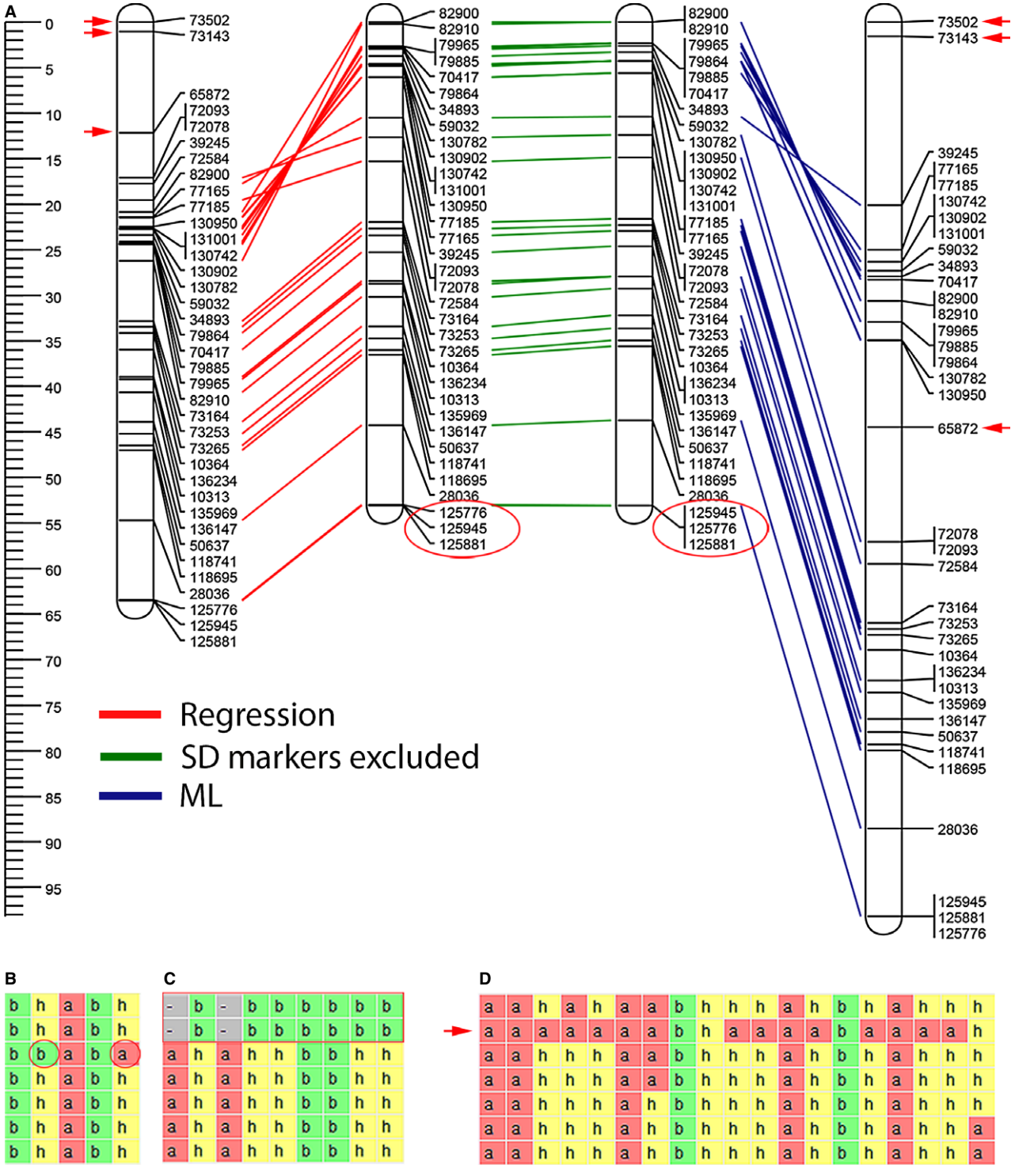
The genus *Melanochromis* (Lake Malawi) would be an interesting group to focus on because it includes species such as *M. auratus* with dark stripes on a lighter background (similar to the focal species of the present study) but also species that have light stripes (e.g. *M. vermivorus*). In addition to the few species in the

Table 3 Empirical estimates of error rates with varying coverage thresholds for accepting markers. The reported values are percentages of the total number of genotypic observations. Individual genotypes that do not meet the coverage threshold are exported as missing genotypic observations

Coverage threshold	8	15	20	30
Number of markers	2219	1686	1517	1253
Missing data	15.05%	19.88%	24.10%	29.47%
False homozygotes	4.09%	3.20%	2.88%	2.26%
False heterozygotes	0.32%	0.18%	0.17%	0.12%

Fig. 4 Effect of genotyping errors in linkage map estimation. (A) The right- and left-most LGs were estimated including markers under severe segregation distortion (shown with red arrows) using the regression (leftmost) and mid-lateral (ML) (rightmost) algorithms. The ruler on the left shows the genetic distances in cM. The lines connect homologous markers. The two LGs in the centre were estimated after removing these SD loci. As can be seen, the SD loci (i) are mostly located at the extremes of the LGs, (ii) are ordered differently depending on the algorithm used, (iii) inflate map distance and (iii) disturb the ordering of other markers. Both the inferred distances and orders are nearly identical after excluding these markers. Note, however, that the markers outlined with red ellipses are completely linked in the ML map, whereas an order and a small genetic distance were inferred using regression. Other examples can be seen throughout this map. Visual inspection revealed that no recombination event was detected between these markers, and therefore, the distance and order estimated inferred using regression is an artefact. Each row represents a marker, and diploid genotypes for individual chromosomes are represented by columns in B–D. (B) The genotypes circled in red are most likely the result of sequencing a single allele. (C) Two markers under SD were genotyped as parental 2 homozygotes (coded as 'b') in >80 F2s prepped in the same batch. The presence of missing data values every time the closest markers is genotyped as a parent 1 homozygote (coded as 'a') suggests systematic allele dropout. (D) The marker shown with the arrow also displayed highly significant SD, and its removal decreased the LG size in 11.5 cM.

Lake Malawi and Lake Victoria radiations, several genera of Lake Tanganyika cichlids have several species that have lateral stripes (e.g. *Chalinochromis*, *Julidochromis*, *Lamprologus*, *Lepidolamprologus* and *Telmatochromis*). The Lake Tanganyika radiation is the oldest of the African radiations and has a more complex history of repeated colonization events and to a certain extent, more independent lineages (Salzburger *et al*. 2005; Koblmuller *et al*. 2008; Takahashi & Koblmüller 2011). In particular, the genus *Julidochromis* contains species that vary both in the number of lateral stripes and also in the degree of contrast to the body coloration. *Julidochr-*



omis regani, for instance, has an additional lateral stripe that is more ventral to the ML. Variation in lateral stripe patterns can also be found in the Neotropical genus of cichlids *Crenicichla* (Kullander & de Lucena 2006). Considering the great divergence time between the African and Neotropical cichlid lineages, it is probable that horizontal stripes have evolved independently in these two lineages. Comparative forward-genetic studies could get at the question of whether parallel evolution at the molecular level has taken place (e.g. if the same mutation, gene or pathway is genetically causal) (Elmer & Meyer 2011; Kronforst *et al.* 2012).

Stripes and bars of cichlid fishes are many times transient indicators of motivational status (Baerends *et al.* 1986; Muske & Fernald 1987; Korzan *et al.* 2008). Interestingly, a number of other cichlids have transient horizontal stripes that are very similar in terms of contrast and position to the ML and DL stripes of *H. sauvagei*. A behavioural study on the riverine species *Chromidotilapia guntheri* reported that lateral stripes are positively correlated with agonistic tendencies (aggression and fleeing) and negatively associated with courtship activity. Specifically, lateral stripes are not associated with aggression, but rather more conspicuous during fleeing and escape (Baerends *et al.* 1986). This finding can probably be extended to most cichlids in general. Several species (e.g. *Pseudocrenilabus multicolor, Astatotilapia burtoni*), particularly ones that are shy (e.g., *H. chilotes*), have visible horizontal stripes when stressed and attempting to flee (FH, personal observation).

It is conceivable therefore that the horizontal stripes in *H. sauvagei* evolved through the loss of regulation of motile activity of the dermal chromatophores. This would also be compatible with the recessive inheritance that we found if a single copy of the ancestral allele is sufficient to ensure motility. Great progress has been made in elucidating the neurological, hormonal and gene expression changes associated with motivational shift in cichlids (Fernald 1976; Muske & Fernald 1987; Parikh *et al.* 2006; Korzan *et al.* 2008; Maan & Sefc 2013). Further characterization and identification of the causal locus we mapped in the present study could help to further understand the molecular genetic mechanisms behind the integration of motivational status and coloration.

RAD-seq data for genetic mapping in nonmodel organisms

Uncritical linkage mapping with RAD-seq data generates an intolerable amount of genotyping errors, erroneous linkages and orders. When numerous SD markers and missing genotype values are present, the process of grouping markers into linkage groups (prior to ordering) requires arbitrary inflation of the LOD cut-off to avoid false groupings. RAD-seq linkage maps tend to be very dense, and problems due to genotyping inaccuracy are aggravated with increasing marker density (Feakes *et al.* 1999; Jansen *et al.* 2001; Hackett & Broadfoot 2003). Although these sources of genotyping error (e.g. allele dropout) also affect the previously used markers types (e.g. microsatellites) (Pompanon *et al.* 2005), errors in NGS data sets are more likely to go undetected due to the sheer number of markers which makes regenotyping and even visual inspection unfeasible for some data sets.

Allelic dropout. An important and probably the most common cause of genotyping errors is allelic dropout (Pompanon *et al.* 2005). This is when one allele fails to be genotyped in a heterozygote. This leads to the overestimation of inbreeding coefficients in population genetic studies (Bonin *et al.* 2004). In linkage mapping, it can inflate distances, lead to wrong ordering and introduce double recombination events. In genotyping-by-sequencing approaches, such dropout can occur simply by chance if one of the alleles is not sequenced. This can be visually detected as it will appear as a single (false) homozygous genotype which is flanked by heterozygous ones (Fig. 4B). The effects of this are not so drastic, and it probably will only lead to a slight inflation of the genetic distance. However, if this occurs in a particularly dense region (where no legitimate recombination event has been observed), it can lead to incorrect ordering. Allelic dropout can also occur systematically and lead to SD when, due to issues such as quality or quantity of DNA (Pompanon *et al.* 2005), several individuals are incorrectly genotyped as homozygous. In addition, allelic dropout can affect markers differently which might result in multiple loci having the same artefactual patterns and therefore being incorrectly linked (Soulsbury *et al.* 2007).

Segregation distortion. In addition to differential susceptibility to errors across markers (e.g. systematic allelic dropout), SD can result from improper mapping parameters and criteria used to consider a marker mappable or informative. For instance, a nonvariable position that is affected by errors might be considered a segregating locus and will lead to higher than expected frequencies of homozygotes in the F2s. The opposite situation, that is higher than expected frequency of heterozygotes, can occur by the merging of paralogous regions and treating them as a single marker (Frisch *et al.* 2004). Errors are treated as recombination by analysis software, and the most notable effect of artefactual SD loci is to inflate map distances. This can, for instance, lead to the incor-

rect placement of these markers at the extremes of linkage groups. Furthermore, because there is a low number of states (a, h and b in a F2 map), systematic errors can easily lead to false groupings and end-to-end joining of LGs which have SD loci in the extremes. Without excluding SD loci, one frequently obtains fewer groups than chromosome numbers, even at extremely high LOD cut-offs of 30, thus indicating spurious linkage (van Ooijen & Jansen 2013). In the present data set, some SD loci led to big inversions (e.g. 15 cM) and severally distorted the order and position of other markers within the same linkage group (Fig. 4).

Segregation distortion can also emerge as a consequence of biological factors, such as sex determination (Kozielska *et al.* 2010), meiotic drive or epistatically lethal recessives between species (e.g. DM incompatibility loci) (Lyttle 1991; Phadnis & Orr 2009). Many mapping projects involve interspecific crosses, and discarding all loci with evidence of SD might impede the identification of interesting cases of SD. It is not trivial to distinguish SD that arises because of systematic errors or because of biological processes. Because of the potential problems and extreme consequences, some have argued to simply remove all loci exhibiting SD above a certain level (van Ooijen & Jansen 2013), but if for whatever reason these cannot be excluded, a few guidelines that can aid in identifying errors are as follows: Does the presence of SD loci affect other markers? Does it lead to incongruent results when using alternative analysis software or algorithms? Do SD loci introduce double recombination events? Are SD loci clustered? (If SD is legitimate, one expects the clustering of several loci with similar levels of SD.) Is there a nonrandom distribution of genotypes based on batches that were prepped together?

Cases of apparently legitimate SD were found on LG18 and LG2. These could be distinguished from error because SD did not lead to double recombination and showed a grouping of several loci with similar levels of distortion. Because our F2 mapping panel was derived from an interspecific cross, this opens the exciting possibility that these SD regions correspond to genes underlying sex determination (Kozielska *et al.* 2010) or postzygotic reproductive isolation (Phadnis & Orr 2009). Other compatible explanation is meiotic drive in the F1 hybrids or linkage to sex determinants. Distinguishing these scenarios awaits experimental evidence.

Missing data. The placement of markers with high levels of missing observations has lower statistical support (van Ooijen & Jansen 2013), and incorrect ordering can arise because of missing observations in the vicinity of recombination events. We observed that depending on the analysis method used, an interval and an order are inferred even in the absence of observed recombination events (Fig. S2, Supporting information). The impact of missing data is therefore significant, and the use of more stringent thresholds (<5% or <10%) is recommended.

Variation in the coverage among individuals in a single sequencing pool results from differences in quality and quantity of DNA across samples. Degradation, variation in the efficiency of restriction enzyme digestion or inaccurate quantification will all lead to a higher incidence of missing data and can only be controlled by careful sample preparation. The systematic occurrence of missing results from technical variation in preparing or sequencing DNA pools. Variation in DNA batch quality or quantity, inaccurate size selection of pooled samples (incomplete overlap of fragment size across libraries) or variation in the number of sequencing reads obtained from different lanes leads to the nonrandom distribution of missing data (Fig. S3, Supporting information).

Suggested solutions to the problem of genotyping errors in RAD data sets

At this point, the best strategy is to use stringent coverage (e.g. >15×) and missing data (e.g. <10%) thresholds for importing markers into the linkage map building software. Post-mapping QC follows by identifying and excluding anomalous markers with evidence of systematic errors (nonrandom occurrence of either missing data or genotyping errors). Recent developments of the Stacks pipeline now allow investigators to filter out loci based on SD level.

The main analysis software packages (Cheema & Dicks 2009) offer innumerous diagnostic tools to spot potential problems in the data set, and these tend to be discussed in depth in the software manuals and other publications (van Ooijen & Jansen 2013). Graphical genotypes can be visually inspected using a variety of software tools including JOINMAP (Van Ooijen 2006) and the freely available R/qtl (Broman & Sen 2009), GGT2 (van Berloo 2008) and Flapjack (Milne *et al.* 2010). An imputation and correction algorithm has been developed to deal with the problems of missing data and nonsystematic allelic dropout in RAD-seq data sets using a backcross design (Ward *et al.* 2013). However, it is important to note that uncritical imputation might lead to strong support for an incorrect order.

To control the incidence of nonsystematic allelic dropout, the average coverage threshold must be determined taking into consideration the probability of sequencing a single allele from a heterozygote. A coverage threshold of at least 15 results in the probability of <0.01% of nonsystematic dropout and would result

in approximately five errors in a F2 data set of 1000 markers and 200 F2s. In our data set, the occurrence of errors was much higher than what would be expected simply by the probability of sequencing a single allele before QC. The proportion of false homozygotes dropped to the expected value (<0.1%) after both QC steps. This indicates that some markers (those excluded in post-mapping QC due to SD) are more prone to errors.

The thresholds for SD and missing data should be determined carefully based on the study's requirements in terms of mapping accuracy and the implications of errors, for example a map developed for Mendelian trait mapping might be considered more tolerant to errors than a map developed to map QTL for complex traits or to scaffold a genome. Furthermore, the downstream analysis can be made more efficient by then excluding markers that are redundant (and have higher levels of missing observations). If a particular region of the map is lacking markers, a reasonable approach is to construct a reliable backbone and to subsequently add markers that map to the region of interest. This was the approach used in the present manuscript to identify four additional markers that mapped within the candidate interval (Fig. 3B).

Although these quality control steps will invariably lead to less dense maps, we stress that this should not be a major concern. The number of markers is rarely the factor limiting resolution in genetic mapping projects (Doerge 2002). If the number of recombination events is not sufficient for resolving the markers, it will lead to highly saturated maps where many markers will be redundant. In the present study, 867 loci already led to a high degree of saturation and redundancy in the present linkage map (Figure S3, Supporting information). Furthermore, the number of possible orders increases exponentially with the number of markers. Therefore, saturated maps with redundant markers pose a significant and unnecessary computational effort. Further resolution can only be obtained by increasing the number of individuals (and observed recombination events) in the mapping panel. The optimal marker density is naturally dependent on the objectives of the study. If the aim is to map genomic regions controlling complex traits, insufficient marker density can lead to the identification of 'ghost QTL', when using interval mapping. Therefore, a median marker distance of 5 cM is recommended (Broman & Speed 1999). If the aim is to do genomic scaffolding, one might wish to anchor the greatest number of scaffolds. But, it must be noted that it will not be possible to determine the orientation and in some cases the order of scaffolds in saturated regions of the map.

Concluding remarks

We show that ML and DL stripes in *H. sauvagei* are (i) under the same genetic control, (ii) are inherited recessively and (iii) map to a 500-kb interval in the tilapia LG18. Our results point to the existence of genetic constraints to the evolution of patterns involving DL and ML stripes in cichlids. A few species have mitigated the constraint posed by the coinheritance of both stripes, and this presumably occurred by the recruitment of genomic regions or variants other than the one identified in the present study. Our findings provide a further example of an adaptive phenotype with a simple genetic basis. Despite the inherent limitations of genetic mapping projects and ascertainment bias aptly pointed out by Rockman (2012), these results set the ground work for future analysis on the molecular mechanisms and evolutionary history that underlies this phenotype.

Furthermore, we show that uncritical RAD-seq generates intolerable amounts of incorrect genotypes that disturb all aspects of linkage map construction and urge investigators to shift focus from the quantity to the quality of genetic markers. With appropriate QC steps, a high genotyping accuracy and appropriate marker density can be obtained using the ddRAD-seq approach and pre-existing software tools, thus enabling effective genetic mapping in nonmodel organisms.

Acknowledgements

This work was funded through a Deutsche Forschungsgemeinschaft grant (DFG ME 1725/11) to AM. We thank the University of Konstanz for supporting the Genomics Center of the University of Konstanz (GeCKo) and the Meyer laboratory. The initial crosses were set up by Dr. Matthias Sanetra and Dr. Helen Gunther.

References

Albertson R, Streelman J, Kocher T (2003) Directional selection has shaped the oral jaws of Lake Malawi cichlid fishes. *Proceedings of the National Academy of Sciences of the United States of America*, **100**, 5252–5257.

Amores A, Catchen J, Ferrara A, Fontenot Q, Postlethwait JH (2011) Genome evolution and meiotic maps by massively parallel DNA sequencing: spotted Gar, an outgroup for the Teleost genome duplication. *Genetics*, **188**, 799–808.

Baerends GP, Wanders JBW, Vodegel R (1986) The Relationship between marking patterns and motivational state in the prespawning behavior of the cichlid fish *Chromidotilapia guentheri* (Sauvage). *Netherlands Journal of Zoology*, **36**, 88–116.

Baird NA, Etter PD, Atwood TS *et al.* (2008) Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS ONE*, **3**, e3376.

Brawand D, Wagner CE, Li YI *et al.* (2014) The genomic substrate for adaptive radiation: genomes of five African cichlid fish. *Nature*, **513**, 375–381.

van Berloo R (2008) GGT 2.0: versatile software for visualization and analysis of genetic data. *Journal of Heredity*, **99**, 232–236.

Bernatchez L, Renaut S, Whiteley AR *et al.* (2010) On the origin of species: insights from the ecological genomics of lake whitefish. *Philosophical Transactions of the Royal Society B-Biological Sciences*, **365**, 1783–1800.

Bonin A, Bellemain E, Eidesen PB *et al.* (2004) How to track and assess genotyping errors in population genetics studies. *Molecular Ecology*, **13**, 3261–3273.

Broman KW, Sen S (2009) *A Guide to QTL Mapping with R/qtl*. Springer, Dordrecht.

Broman KW, Speed TP (1999) A review of methods for identifying QTLs in experimental crosses. In: Statistics in molecular biology and genetics: selected proceedings of a 1997 Joint AMS-IMS-SIAM summer conference on statistics in molecular biology (ed. Seillier-Moiseiwitsch Fo), pp. vi, 313 p. American Mathematical Society, Providence, R.I.Hayward, Calif.

Carleton KL, Hofmann CM, Klisz C *et al.* (2010) Genetic basis of differential opsin gene expression in cichlid fishes. *Journal of Evolutionary Biology*, **23**, 840–853.

Catchen JM, Amores A, Hohenlohe P, Cresko W, Postlethwait JH (2011) Stacks: building and genotyping loci de novo from short-read sequences. *G3-Genes Genomes. Genetics*, **1**, 171–182.

Chan YF, Marks ME, Jones FC *et al.* (2010) Adaptive evolution of pelvic reduction in sticklebacks by recurrent deletion of a Pitx1 enhancer. *Science*, **327**, 302–305.

Cheema J, Dicks J (2009) Computational approaches and software tools for genetic linkage map estimation in plants. *Briefings in Bioinformatics*, **10**, 595–608.

Cnaani A, Lee BY, Zilberman N *et al.* (2008) Genetics of sex determination in tilapiine species. *Sexual Development*, **2**, 43–54.

Davey JW, Hohenlohe PA, Etter PD, *et al.* (2011) Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nature Reviews Genetics*, **12**, 499–510.

Doerge RW (2002) Mapping and analysis of quantitative trait loci in experimental populations. *Nature Reviews Genetics*, **3**, 43–52.

Domingues VS, Poh Y-P, Peterson BK *et al.* (2012) Evidence of adaptation from ancestral variation in young populations of beach mice. *Evolution*, **66**, 3209–3223.

Elmer KR, Meyer A (2011) Adaptation in the age of ecological genomics: insights from parallelism and convergence. *Trends in Ecology & Evolution*, **26**, 298–306.

Elmer KR, Reggio C, Wirth T *et al.* (2009) Pleistocene desiccation in East Africa bottlenecked but did not extirpate the adaptive radiation of Lake Victoria haplochromine cichlid fishes. *Proceedings of the National Academy of Sciences of the United States of America*, **106**, 13404–13409.

Feakes R, Sawcer S, Chataway J *et al.* (1999) Exploring the dense mapping of a region of potential linkage in complex disease: an example in multiple sclerosis. *Genetic Epidemiology*, **17**, 51–63.

Fernald RD (1976) Effect of testosterone on behavior and coloration of adult male cichlid fish (*Haplochromis burtoni*, Gunther). *Hormone Research*, **7**, 172–178.

Franchini P, Fruciano C, Spreitzer ML *et al.* (2013) Genomic architecture of ecologically divergent body shape in a pair of sympatric crater lake cichlid fishes. *Molecular Ecology*, **23**, 1828–1845.

Frisch M, Quint M, Lubberstedt T, Melchinger AE (2004) Duplicate marker loci can result in incorrect locus orders on linkage maps. *Theoretical and Applied Genetics*, **109**, 305–316.

Hackett CA, Broadfoot LB (2003) Effects of genotyping errors, missing values and segregation distortion in molecular marker data on the construction of linkage maps. *Heredity*, **90**, 33–38.

Henning F, Meyer A (2014) The evolutionary genomics of cichlid fishes: explosive speciation and adaptation in the postgenomic era. *Annual Reviews of Genomics and Human Genetics*, **15**, 417–441.

Hoekstra HE, Hirschmann RJ, Bundey RA, Insel PA, Crossland JP (2006) A single amino acid mutation contributes to adaptive beach mouse color pattern. *Science*, **313**, 101–104.

Jansen J, de Jong AG, van Ooijen JW (2001) Constructing dense genetic linkage maps. *Theoretical and Applied Genetics*, **102**, 1113–1122.

Johnston SE, Gratten J, Berenos C *et al.* (2013) Life history trade-offs at a single locus maintain sexually selected genetic variation. *Nature*, **502**, 93–95.

Jones FC, Grabherr MG, Chan YF *et al.* (2012) The genomic basis of adaptive evolution in threespine sticklebacks. *Nature*, **484**, 55–61.

Kawakami K, Amsterdam A, Shimoda N *et al.* (2000) Proviral insertions in the zebrafish hagoromo gene, encoding an F-box/WD40-repeat protein, cause stripe pattern anomalies. *Current Biology*, **10**, 463–466.

Koblmuller S, Sefc KM, Sturmbauer C (2008) The Lake Tanganyika cichlid species assemblage: recent advances in molecular phylogenetics. *Hydrobiologia*, **615**, 5–20.

Kocher T (2004) Adaptive evolution and explosive speciation: the cichlid fish model. *Nature Reviews Genetics*, **5**, 288–298.

Kocher T, Lee W-J, Sobolewska H, Penman D, McAndrew B (1998) A genetic linkage map of the cichlid fish, the tilapia (*Oreochromis niloticus*). *Genetics*, **148**, 1225–1232.

Korzan WJ, Robison RRB, Zhao S, Fernald RD (2008) Color change as a potential behavioral strategy. *Hormones and Behavior*, **54**, 463–470.

Kozielska M, Weissing FJ, Beukeboom LW, Pen I (2010) Segregation distortion and the evolution of sex-determining mechanisms. *Heredity*, **104**, 100–112.

Kronforst MR, Barsh GS, Kopp A *et al.* (2012) Unraveling the thread of nature's tapestry: the genetics of diversity and convergence in animal pigmentation. *Pigment Cell and Melanoma Research*, **25**, 411–433.

Kullander SO, de Lucena CAS (2006) A review of the species of *Crenicichla* (Teleostei: Cichlidae) from the Atlantic coastal rivers of southeastern Brazil from Bahia to Rio Grande do Sul States, with descriptions of three new species. *Neotropical Ichthyology*, **4**, 127–146.

Langmead B, Trapnell C, Pop M, Salzberg SL (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology*, **10**, R25.

Lee B-Y, Lee W-J, Streelman J *et al.* (2005) A second-generation genetic linkage map of tilapia (*Oreochromis* spp.) *Genetics*, **170**, 237–244.

Linnen CR, Kingsley EP, Jensen JD, Hoekstra HE (2009) On the origin and spread of an adaptive allele in deer mice. *Science*, **325**, 1095–1098.

Loh Y-HE, Bezault E, Muenzel FM *et al.* (2013) Origins of shared genetic variation in African cichlids. *Molecular Biology and Evolution*, **30**, 906–917.

Lyttle TW (1991) Segregation distorters. *Annual Review of Genetics*, **25**, 511–557.

Maan ME, Sefc KM (2013) Colour variation in cichlid fish: developmental mechanisms, selective pressures and evolutionary consequences. *Seminars in Cell & Developmental Biology*, **24**, 516–528.

Maan ME, Seehausen O, Soderberg L *et al.* (2004) Intraspecific sexual selection on a speciation trait, male coloration, in the Lake Victoria cichlid *Pundamilia nyererei*. *Proceedings of the Royal Society of London Series B-Biological Sciences*, **271**, 2445–2452.

Manceau M, Domingues VS, Linnen CR, Rosenblum EB, Hoekstra HE (2010) Convergence in pigmentation at multiple levels: mutations, genes and function. *Philosophical Transactions of the Royal Society B-Biological Sciences*, **365**, 2439–2450.

Mazzuchelli J, Kocher TD, Yang FT, Martins C (2012) Integrating cytogenetics and genomics in comparative evolutionary studies of cichlid fish. *BMC Genomics*, **13**, 463.

Meyer A, Biermann CH, Orti G (1993) The phylogenetic position of the zebrafish (*Danio rerio*), a model system in developmental biology: an invitation to the comparative method. *Proceedings of Biological Sciences*, **252**, 231–236.

Miller MR, Dunham JP, Amores A, Cresko WA, Johnson EA (2007) Rapid and cost-effective polymorphism identification and genotyping using restriction site associated DNA (RAD) markers. *Genome Research*, **17**, 240–248.

Milne I, Shaw P, Stephen G *et al.* (2010) Flapjack-graphical genotype visualization. *Bioinformatics*, **26**, 3133–3134.

Muske LE, Fernald RD (1987) Control of a teleost social signal. *Journal of Comparative Physiology A: Sensory, Neural, and Behavioral Physiology*, **160**, 99–107.

Nadeau NJ, Jiggins CD (2010) A golden age for evolutionary genetics? Genomic studies of adaptation in natural populations. *Trends in Genetics*, **26**, 484–492.

van Ooijen JW, Jansen J (2013) *Genetic Mapping in Experimental Populations*. Cambridge University Press, Cambridge.

O'Quin CT, Drilea AC, Roberts RB, Kocher TD (2012) A small number of genes underlie male pigmentation traits in Lake Malawi cichlid fishes. *Journal of Experimental Zoology Part B-Molecular and Developmental Evolution*, **318B**, 199–208.

O'Quin CT, Drilea AC, Conte MA, Kocher TD (2013) Mapping of pigmentation QTL on an anchored genome assembly of the cichlid fish, *Metriaclima zebra*. *BMC Genomics*, **14**, 287.

Orr HA (2005) The genetic theory of adaptation: a brief history. *Nature Reviews Genetics*, **6**, 119–127.

Parikh VN, Clement TS, Fernald RD (2006) Androgen level and male social status in the African cichlid, Astatotilapia burtoni. *Behavioural Brain Research*, **166**, 291–295.

Peterson BK, Weber JN, Kay EH, Fisher HS, Hoekstra HE (2012) Double digest RADseq: an inexpensive method for *de novo* SNP discovery and genotyping in model and non-model species. *PLoS ONE*, **7**, e37135.

Phadnis N, Orr HA (2009) A single gene causes both male sterility and segregation distortion in *Drosophila* hybrids. *Science*, **323**, 376–379.

Pompanon F, Bonin A, Bellemain E, Taberlet P (2005) Genotyping errors: causes, consequences and solutions. *Nature Reviews Genetics*, **6**, 847–859.

Recknagel H, Elmer KR, Meyer A (2013) A hybrid genetic linkage map of two ecologically and morphologically divergent Midas cichlid fishes (*Amphilophus* spp.) obtained by massively parallel DNA sequencing (ddRADSeq). *G3-Genes Genomes. Genetics*, **3**, 65–74.

Reed RD, Papa R, Martin A *et al.* (2011) optix drives the repeated convergent evolution of butterfly wing pattern mimicry. *Science*, **333**, 1137–1141.

Rockman MV (2012) The QTN program and the alleles that matter for evolution: all that's gold does not glitter. *Evolution*, **66**, 1–17.

Roesti M, Moser D, Berner D (2013) Recombination in the threespine stickleback genome-patterns and consequences. *Molecular Ecology*, **22**, 3014–3027.

Salzburger W, Mack T, Verheyen E, Meyer A (2005) Out of Tanganyika: genesis, explosive speciation, key-innovations and phylogeography of the haplochromine cichlid fishes. *BMC Evolutionary Biology*, **5**, 17.

Sanetra M, Henning F, Fukamachi S, Meyer A (2009) A microsatellite-based genetic linkage map of the cichlid fish, *Astatotilapia burtoni* (Teleostei): a comparison of genetic architectures among rapidly speciating cichlids. *Genetics*, **182**, 387–397.

Schluter D (1996) Adaptive radiation along genetic lines of least resistance. *Evolution*, **50**, 1766–1774.

Seehausen O (2000) Explosive speciation rates and unusual species richness in haplochromine cichlid fishes: effects of sexual selection. *Advances in Ecological Research*, **31**, 237–274.

Seehausen O, Mayhew PJ, Van Alphen JJM (1999) Evolution of colour patterns in East African cichlid fish. *Journal of Evolutionary Biology*, **12**, 514–534.

Slate J (2013) 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.

Soulsbury CD, Iossa G, Edwards KJ, Baker PJ, Harris S (2007) Allelic dropout from a high-quality DNA source. *Conservation Genetics*, **8**, 733–738.

Stinchcombe JR, Hoekstra HE (2008) Combining population genomics and quantitative genetics: finding the genes underlying ecologically important traits. *Heredity*, **100**, 158–170.

Stirnimann CU, Petsalaki E, Russell RB, Muller CW (2010) WD40 proteins propel cellular networks. *Trends in Biochemical Sciences*, **35**, 565–574.

Streelman JT, Kocher TD (2000) From phenotype to genotype. *Evolution and Development*, **2**, 166–173.

Streelman JT, Albertson RC, Kocher TD (2003) Genome mapping of the orange blotch colour pattern in cichlid fishes. *Molecular Ecology*, **12**, 2465–2471.

Takahashi T, Koblmüller S (2011) The adaptive radiation of cichlid fish in Lake Tanganyika: a morphological perspective. *International Journal of Evolutionary Biology*, **2011**, 620754.

Terai Y, Morikawa N, Kawakami K, Okada N (2002) Accelerated evolution of the surface amino acids in the WD-repeat domain encoded by the hagoromo gene in an explosively speciated lineage of east African cichlid fishes. *Molecular Biology and Evolution*, **19**, 574–578.

Van Ooijen JW (2006) JoinMap 4. Software for the calculation of genetic linkage maps in experimental populations. Kyazma BV, Wageningen, The Netherlands.

Verheyen E, Salzburger W, Snoeks J, Meyer A (2003) Origin of the superflock of cichlid fishes from Lake Victoria, East Africa. *Science*, **300**, 325–329.

Wang S, Basten CJ, Zeng Z-B (2012) *Windows QTL Cartographer 2.5*. Department of Statistics, North Carolina State University, Raleigh, North Carolina. Available at: http://statgen.ncsu.edu/qtlcart/WQTLCart.htm.

Ward JA, Bhangoo J, Fernandez-Fernandez F *et al*. (2013) Saturated linkage map construction in *Rubus idaeus* using genotyping by sequencing and genome-independent imputation. *BMC Genomics*, **14**, 2.

A.M., H.L. and F.H. designed the study. Fish breeding and phenotyping were carried out by H.L. Molecular analyses were performed under the supervision of P.F. at the Genomics Center of the University of Konstanz (GeCKo). F.H. and P.F. analysed the data. F.H. drafted the manuscript. All authors edited and agreed to the manuscript.

Data accessibility

Raw Illumina sequences were deposited into the NCBI's Sequence Read Archive (SRA) database with Accession no. SRA171565. The final linkage map, both JoinMap v.4 input files used for map construction and fine mapping, and the phenotypic data were deposited into the DRYAD database (doi:10.5061/dryad.r1f52).

Supporting information

Additional supporting information may be found in the online version of this article.

Fig. S1 Histogram of the inferred genetic distances between adjacent markers.

Fig. S2 Genetic distances inferred in the absence of recombination events.

Fig. S3 Systematic occurrence of missing data due to library preparation effects.

Table S1 Sequencing statistics of the five RAD libraries.

Table S2 Genotypic frequencies for the 867 loci present in the final linkage map.

Table S3 Individual genotypic frequencies.