Erschienen in: Molecular Ecology ; 26 (2017), 1. - S. 77-91
https://dx.doi.org/10.1111/mec.13699

Towards understanding the genetic basis of mouth asymmetry in the scale-eating cichlid *Perissodus microlepis*

FRANCESCA RAFFINI,*† CARMELO FRUCIANO,*[1] PAOLO FRANCHINI* and AXEL MEYER*†
**Lehrstuhl für Zoologie und Evolutionsbiologie, Department of Biology, University of Konstanz, Universitätsstrasse 10, 78464 Konstanz, Germany, †International Max Planck Research School (IMPRS) for Organismal Biology, Max-Planck-Institut für Ornithologie, Am Obstberg 1, 78315 Radolfzell, Germany*

Abstract

How polymorphisms consisting in left right asymmetries are produced and maintained in natural populations is a tantalizing question, which remains largely unanswered. The scale-eating cichlid fish *Perissodus microlepis* is a remarkable example of extreme ecological specialization achieved by morphological and behavioural laterality. Its asymmetric mouth is accompanied by a pronounced lateralized foraging behaviour, where a left-bending morph preferentially feeds on the scales of the right side of its prey, while the opposite is true for the right morph. This striking asymmetry made this fish a textbook example of the astounding degree of ecological specialization and negative frequency-dependent selection. Yet, the genetic basis underlying this spectacular laterality remains unknown. We addressed this question through analyses of wild-caught fish using high-throughput DNA sequencing data. A novel array of SNP markers was developed by ddRAD sequencing (ddRADseq) and the use of pooled DNA samples (PoolSeq). We obtained more than 155 000 SNPs using ddRADseq and 3 900 000 SNPs with PoolSeq. Among these, we identified one (ddRAD) SNP, and 38 or 378 (PoolSeq) windows that are differentiated between the left and right morphs accounting for spurious associations due to geographic structuring. This allowed us to uncover candidate genomic regions that potentially contain genes for this trait. Then, this interesting trait has a genetic basis that is likely to be influenced by multiple loci. This result contributes to a greater understanding of the genetic bases of left right asymmetry and, ultimately, the evolutionary processes governing the maintenance of this striking case of laterality.

Keywords: bilateral asymmetry, *Perissodus microlepis*, PoolSeq, quantitative trait, RADseq, scale-eating cichlid fish

Introduction

Natural selection is a process that results from the differential survival and reproduction of those individuals that are better than others adapted to the prevailing environmental conditions. The survivors tend to produce more offspring than those less well adapted, so that the characteristics of the population change over time, promoting the evolution of adaptive traits (Darwin 1859). Crucial in this process is phenotypic variation, which plays a significant role in the ecology and evolution of natural populations. The distribution of phenotypic values itself can be shaped by natural selection, as clearly shown in the industrial melanism of the peppered moth (reviewed in Cook & Saccheri 2013). Discontinuous phenotypes (such as the melanic and

Correspondence: Axel Meyer, Fax: +49 (0) 7531 88 3018; E mail: axel.meyer@uni konstanz.de
[1]Present address: School of Earth, Environmental & Biological Sciences, Queensland University of Technology, Brisbane, Qld 4000, Australia

typical moth forms) are known as polymorphism (Robinson & Schluter 2000). Another notable example of such a polymorphism is left right asymmetry or bilateral asymmetry, where left and right individuals differ from a typically bilateral symmetrical individual (Palmer 2004). This kind of asymmetry has been found in several groups of animals, for example, in eye side in flatfish (*Pleuronectiformes*; Hubbs & Hubbs 1945), shell coiling direction of tree snails (*Amphidromus* spp.; Sutcharit *et al.* 2007) and direction of the mouth opening in the cichlid fish *Perissodus microlepis* (Hori 1993; Lee *et al.* 2010, 2015; Kusche *et al.* 2012).

Perissodus microlepis is one of the nine species of scale-eating cichlids of the tribe Perissodini endemic to Lake Tanganyika, Africa (Koblmüller *et al.* 2007; Takahashi *et al.* 2007). This fish has received special attention from evolutionary biologists during the last 20 years, and it has become a striking example of the extreme degree of morphological and ecological specialization produced by the adaptive radiation of African cichlids (reviewed in Henning & Meyer 2014; Meyer 2015). Two morphs have been initially described within this species with respect to mouth-opening direction: one morph has the mouth turned to the right ('right' morph) and the other morph's mouth opens towards the left ('left' morph; Hori 1993). This remarkable polymorphism is seen as an extreme case of adaptive evolution (Lee *et al.* 2015), as it is associated with lateralized foraging behaviour. *Perissodus microlepis* is mainly a lepidophagous predator (Nshombo *et al.* 1985; Takeuchi *et al.* 2016), and left morphs preferentially attack the prey's right side, while the opposite applies to the right morph, increasing the hunting success (Hori 1993; Van Dooren *et al.* 2010; Lee *et al.* 2012; Takeuchi *et al.* 2012). However, fitness is a relative measure, and the adaptive value of a morphological trait is not always fixed either, but in some circumstances can vary depending on the abundance of alternative phenotypes that is, frequency-dependent selection. This appears to be the case of *P. microlepis*, whose equal abundance of both morphs observed within populations (Hori 1993; Kusche *et al.* 2012) is considered maintained by the advantage of the less frequent morph (known as negative frequency-dependent selection; Hori 1993; Nakajima *et al.* 2004). A single Mendelian locus with two alleles (L and R, R dominant and homozygous lethal; Hori 1993; Hori *et al.* 2007) and linked to a microsatellite locus (UNH2101; Stewart & Albertson 2010) has been proposed to control mouth asymmetry. A similar relationship between morphology and behaviour, and the same genetic determination mode, was observed also in other fishes exhibiting mouth asymmetry (e.g. Mboko *et al.* 1998; Seki *et al.* 2000; Hori 2000; Hori *et al.* 2007; Nakajima *et al.* 2007; Takeuchi & Hori 2008; Yasugi & Hori 2011; Seki *et al.* 2000; Stewart & Albertson 2010; Hata *et al.* 2012; Hata & Hori 2012). However, mouth asymmetry has been recently found to have a continuous unimodal distribution in *P. microlepis* (Van Dooren *et al.* 2010; Kusche *et al.* 2012), rather than the two clear discrete states originally described (Hori 1993). These findings challenged also the single gene determination model (Hori 1993; Hori *et al.* 2007; Stewart & Albertson 2010), since this mode implies the absence of near-symmetrical samples. Additionally, it has been shown that this genetic model is not consistent with published offspring phenotype frequencies (Palmer 2010; Lee *et al.* 2015), and mouth asymmetry is not associated with the proposed microsatellite locus (Lee *et al.* 2010, 2015). These studies contribute to the mounting evidence that this fascinating textbook model (Futuyma 2009) might not be so simple and clear as initially proposed (Hori 1993; Palmer 2010), and understanding the mechanisms driving the evolution of *P. microlepis* intraspecific diversity is now more intriguing than ever.

Here, we aim to shed light on the genetic basis of this remarkable polymorphism. Clarifying its genetic determination is a crucial step towards understanding the driver(s) of this iconic trait. Several studies have directly or indirectly focused on the genetic basis of mouth asymmetry, but this continues to be elusive. In most occurrences of bilateral asymmetries exhibiting equal abundance of the left and right morphs, the direction of asymmetry is not inherited (27 of 28; Palmer 2004). Consequently, it has been hypothesized that it is purely random and not genetically determined also in *P. microlepis* (random antisymmetry model; Palmer 2005). An experiment in which *P. microlepis* was forced to feed only on one side (Van Dooren *et al.* 2010), observations in both laboratory-reared (Lee *et al.* 2012) and wild-caught (Kusche *et al.* 2012) fish and analysis of stomach contents (Takeuchi *et al.* 2016) all suggest that this trait can be influenced by external factors such as predation mode and feeding experience. These are all elements possibly contributing to the elusiveness of the genetic basis of this trait. On the other hand, observed offspring frequencies did not fit the random model (Palmer 2010). Additionally, reasonable levels of heritability for mouth asymmetry have been described recently (Lee *et al.* 2015), and several lines of evidence including the continuous distribution of the phenotype (Kusche *et al.* 2012) suggest that mouth asymmetry should be a quantitative trait (Kusche *et al.* 2012; Lee *et al.* 2015). Furthermore, gene(s) underlying this trait might not influence mouth asymmetry directly, as previously speculated, but indirectly through their impact on behavioural laterality (Van Dooren *et al.* 2010; Lee *et al.* 2012). Is mouth asymmetry solely environmentally determined (i.e. random), or does it have a sizable

genetic basis? If variation in this trait is (at least partially) genetically determined, is it controlled by a single locus or multiple genomic regions? Is mouth asymmetry driven by behavioural lateralization (Van Dooren *et al.* 2010; Lee *et al.* 2012), or is handed behaviour a consequence of morphological asymmetry (Hori 1993; Takeuchi *et al.* 2016)? To address the genetic basis and to identify candidate region(s) involved in *P. microlepis* mouth asymmetry, we analysed wild-caught specimens using high-throughput DNA sequencing data.

Quantitative trait locus (QTL) mapping analysis represents the approach that has been traditionally used for bridging the gap between phenotypic traits (e.g. mouth asymmetry) and their underlying genes. However, *P. microlepis* fish husbandry is particularly difficult (Lee *et al.* 2010), and, due to the relatively small brood sizes of this species, it proved difficult to obtain enough individuals for QTL mapping. Consequently, we used an alternative approach to identify the genetic bases of mouth asymmetry, based on the comparison of wild-caught samples grouped according to their mouth phenotype. To maximize the power of detecting genomic regions underlying the trait of interest, we analysed only the individuals with the most extreme phenotype (a method commonly used in bulked segregant analysis; Michelmore *et al.* 1991). Additionally, we used two different next-generation sequencing methods. This approach allowed us to obtain a higher number of markers spanning a higher number of different regions than the each technique alone would allow, thus increasing our chances of sequencing genomic regions containing genes underlying mouth asymmetry. Specifically, we developed a novel array of SNP markers *via* (i) individual sequencing through double-digest restriction-associated DNA (RAD) tags (ddRADseq; Miller *et al.* 2007; Baird *et al.* 2008; Peterson *et al.* 2012) and (ii) the sequencing of pooled DNA samples (PoolSeq; Futschik & Schlotterer 2010). These methods allow generating a large amount of SNPs in a quick, efficient and cost-effective manner, and these markers can then be used to uncover the genetic bases of phenotypic traits (Ehrenreich *et al.* 2009; Magwene *et al.* 2011; Kofler *et al.* 2011a). Using these approaches, we aimed to obtain the first empirical information on the genomic architecture of mouth asymmetry in *P. microlepis*, a nonmodel species lacking any previous genomic information.

The genotype phenotype correlations found to be noncausal due to the presence of population structure have been a great concern in uncovering nucleotide variants for complex traits. In fact, differences in allele frequencies between populations due to systematic differences in ancestry (population structuring) rather than association of genes with trait of interest can invalidate the identification of candidate genomic regions, leading to apparent associations at markers that are unlinked to the trait loci (false positives; Pritchard & Donnelly 2001; Freedman *et al.* 2004; Price *et al.* 2006; Balding 2006; Ehrenreich *et al.* 2009; Shin & Lee 2015; Wellenreuther & Hansonn 2016 and references therein). Population structure is influenced both by biological features such as ecological specialization and dispersal potential, as well as external environmental factors such as geography and habitat structure. It is, then, of utmost importance to take into account the patterns of geographic structuring when the analysed samples are differentiated between sampling locations. The phylogeographic structure of *P. microlepis* along the Southern Lake Tanganyika coast (Zambia) has been described in Koblmüller *et al.* (2009) and Lee *et al.* (2010). Their results indicated the presence of significant differentiation also at small spatial scales. Therefore, to limit the occurrence of false-positive candidate SNPs linked to mouth asymmetry, we integrated in our analyses information on geographic provenance by (i) testing for geographic differentiation in our data set and (ii) treating geographic provenance as a confounding factor.

This study represents the first investigation of the genetic basis of mouth asymmetry in *P. microlepis* based on a genome-wide set of a very large number of DNA markers. Our approach allowed to identify SNPs differentiated between *P. microlepis* individuals that are the extreme ends of the distribution of left and right morphs and hence are candidate genomic regions for bilateral asymmetry.

Materials and methods

Sampling and phenotype scoring

Two hundred and sixty six *Perissodus microlepis* adult individuals were collected at seven sites across Lake Tanganyika, three in Congo and four in Zambia (Fig. 1a; Table S1, Supporting information). The samples from Zambia were collected in April 2010 (Kusche *et al.* 2012), while the specimens from Congo were collected in September 2013. Due to the small geographic distance between the three Congo sites (Table S1, Supporting information), and to their small sample sizes, they were considered as a single population. We chose this sampling design to be able to study the genetic basis of mouth asymmetry while controlling for the potentially confounding factor of geographic structure. As not much is known on the genetic basis of mouth asymmetry and this might be different in different populations, we preferred this sampling/analytical design to the alternative sampling of a single population and assuming that the results would generalize to all the populations of the species. Specimens were preserved in

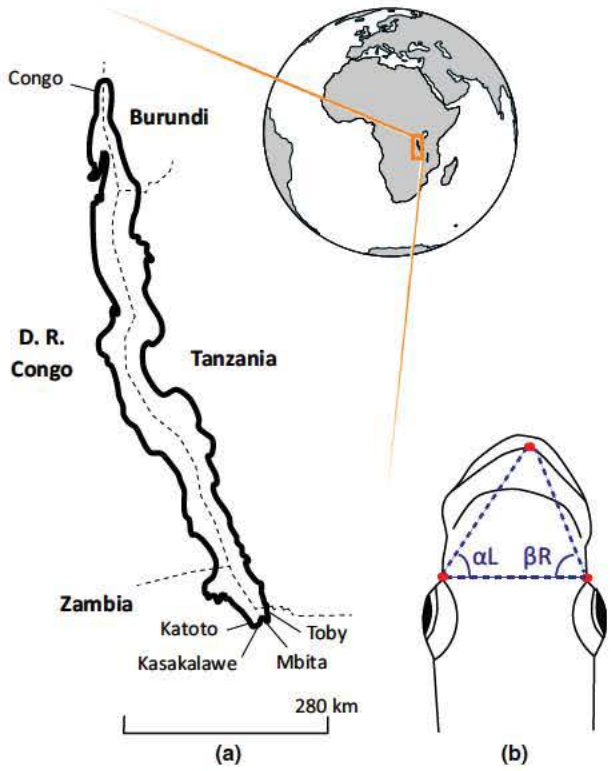

Fig. 1 (a) Lake Tanganyika sampling locations in Zambia and Congo (Africa). Countries are reported in boldface, sampling sites with regular front. (b) Phenotype scoring: the difference between angles at the left (αL) and right (βR) eye measures the degree of laterality of each individual.

ethanol at 4 °C, and finclips were dissected for DNA extraction. Fishes were photographed using procedures aimed to minimize bias and error during data collection (Fruciano *et al.* 2011a,b; Fruciano 2016). For each individual, we recorded, using the software TPSDIG2 v. 2.18 (Rohlf 2006), the x,y coordinates of three points corresponding to the most anterior part of the eye sockets and the tip of the snout, as observed on the upper lip (Fig. 1b). From the coordinates of these points, we computed the angles at each of the eye sockets and used these to determine the mouth-bending angle (a measure of the amount of asymmetry for each individual; Kusche *et al.* 2012). Briefly, the angle αL is the angle formed connecting these three points and having the vertex at the left eye, while the one with the vertex at the right eye is labelled βR (Fig. 1b). The mouth-bending angle was defined as the difference in degrees between the angles at the left and right eye (αL βR). Positive values indicate left-bending individuals, whereas negative results are right-bending fish (Kusche *et al.* 2012). To ensure accurate measurements, we performed a preliminary analysis of measurement error by taking for a pilot set of 20 specimens repeated measurements (two pictures and two digitization per picture, for a total of four measurements; Fruciano *et al.* 2011a,b; 2012) and measuring the consistency of the mouth-bending angle across repeated measurements (repeatability) with the intraclass correlation coefficient (Fisher 1958; Fleiss & Shrout 1977). The value of the repeatability of mouth-bending angle was high (0.89) and for the rest of the data set a single measurement was deemed sufficiently accurate (Fruciano 2016).

Sample selection

Individuals were ranked based on the measured angles, and 50 samples from both tails of the phenotypic distribution were selected, creating two groups of the 25 most extreme right and 25 most extreme left fishes, equally distributed between the five sampling location (Table S1, Supporting information). This sample size has been proven to be large enough to screen candidate markers (Wang *et al.* 2014 and references therein). For the PoolSeq data set, the number of individuals for each morph was increased to 50 (Schlötterer *et al.* 2014). These were evenly distributed between the four Zambian sampling sites obtaining four pools for each morph (Table S1, Supporting information). The samples from Congo, while used for the ddRADseq analyses, were excluded from the PoolSeq analyses due to the low sample size of this population. We focused on two sets of analyses: differentiation between left and right morphs (genetic bases of mouth asymmetry; henceforth 'morph data set'), and among the five sampling sites (geographic structuring; 'geographic data set'). The latter has been used to test for the need of controlling for geographic structuring when analysing the morph data set.

Molecular methods

Genomic DNA was extracted from fin tissue using the ZR Genomic DNATM-Tissue MiniPrep kit (Zymo Research, Irvine, CA, USA) following the manufacturer's protocol including the RNase treatment to remove residual RNA. The DNA integrity of each sample was assessed by agarose gel electrophoresis and quantified using a QUBIT v2.0 fluorometer (Life Technologies, Darmstadt, Germany). Approximately 700 ng of DNA template of each sample was double-digested using the restriction enzymes *PstI-HF* and *MspI* (New England BioLabs, Beverly, MA, USA) in one combined reaction as described in Franchini *et al.* (2014). The library was size-selected for a range of 350 490 bp using a Pippin Prep electrophoresis system (Sage Science, Beverly, MA, USA).

Two and a half micro gram of pooled DNA was used to prepare the PoolSeq library following the Illumina TruSeq DNA Sample Preparation Kit protocol (Illumina Inc., San Diego, CA, USA). The size was selected to 400 600 bp using the Pippin Prep system.

The ddRAD and PoolSeq libraries were individually run on an Illumina HiSeq 2500 (two lanes in total) at the Tufts University Genomics Center (TUCF Genomics, Boston, MA, USA) using the single-end (ddRAD, 151 cycles) and paired-end (PoolSeq, 302 cycles) strategies.

ddRAD bioinformatic pipelines

Raw ddRAD Illumina reads were processed into candidate RAD loci using the *process radtags* script implemented in the STACKS PIPELINE v. 1.28 (Catchen *et al.* 2013). Sequences of each individual were grouped by barcode and quality controlled (final length 146 bp). The filtered reads were *de novo* assembled through the Stacks *denovo map.pl* script, using the following parameters: minimum stack depth (-m) 3, distance allowed between catalogue loci (-n) 3 and removal of highly repetitive RAD tags (-t). This data set was corrected using the Stacks *rxstacks* script and the following settings: prune out haplotypes unlikely to occur in the population (prune haplo), SNP bounded model (--model type bounded), epsilon upper bound (--bound high) 0.1, filter catalogue loci having a log likelihood lower than (--lnl - filter --lnl limx) -10, filter confounding loci (--conf filter), proportion of confounding loci (--conf lim) 0.25. For the analysis of geographic structuring, we tested each locus for deviations from Hardy Weinberg equilibrium (HWE) in each population separately using PLINK v. 1.9 (Purcell *et al.* 2007) and excluding ('blacklisted' in Stacks) from subsequent analyses those loci showing a significant departures from HWE. This procedure allowed us to filter out those loci potentially linked to other evolutionary processes that might confound the signature of geographic differentiation (Wigginton *et al.* 2005). Since marker trait association accompanied by selection can lead to deviations of the HWE (Wigginton *et al.* 2005), the HWE filtering was not applied in the comparison between morphs.

The left and right groups, as well as the five geographic sites, were compared at each locus through pairwise F_{ST} (Weir & Cockerham 1984; Nielsen & Beaumont 2009) and the Fisher's exact test (Fisher 1958) as implemented in the Stacks *populations* module. The minimum percentage of individuals in a population required to process a locus for that population (-r) was set at 0.4, together with 5 individual minimum stack depth required for individuals at a locus (-m). The *P*-values were corrected for multiple tests in SGOF+ v. 3.8 (Carvajal-Rodriguez & de Uña-Alvarez 2011). This software implements multiple correction methods, and we used both the Benjamini & Hochberg (1995; BH hereafter) and the sequential Bonferroni (Holm 1979; SB hereafter) procedure to include approaches based on different philosophies and having different levels of power. SNPs significantly differentiated in both the comparison between morphs and between sites were excluded from the morph data set to reduce the chance of false positive due to population structuration. A Manhattan plot of the F_{ST} values between the left and right fish was obtained using the R package *qqman* (Turner 2014). The position of each SNP was inferred by blasting on the *Oreochromis niloticus* genome, the only anchored reference genome available for cichlids (Brawand *et al.* 2014). When SNPs did not blast on this genome, the *Maylandia zebra* (Brawand *et al.* 2014) one was used as reference genome. To ensure the robustness of the SNPs detected as differentiated between the left and right group, the *de novo* assembly procedure was repeated for the morph comparison excluding samples from Congo, or using default settings, and multiple values and combinations of the following parameters: minimum depth of coverage required to create a stack (*ustacks* m: 2, 3, 5, 10), maximum distance allowed between stacks (*ustacks* -M: 3, 5), maximum number of stacks at a single *de novo* locus (*ustacks* --max locus stacks: 2, 6), number of mismatches allowed between sample tags when generating the catalogue (*cstacks* -n: 2, 3, 5, 10) and upper bound for the error rate (*rxstacks* --bound high*:* 0.05, 0.1).

Genetic relationship between the geographic sites has been further analysed through the principal component analysis (PCA) using the R v. 3.2.0 (R core team 2013) library ADEGENET v. 1.4 2 (Jombart & Ahmed 2011).

To control for the influence of geographic structuring on the analysis of differentiation between morphs, allele frequencies of two types of data sets were subjected to hierarchical analyses of molecular variance (AMOVA; Excoffier *et al.* 1992) in ARLEQUIN v. 3.5 (Excoffier & Lischer 2010). We modelled genetic variation as a function of a morph (main term) and geographic provenance (term of the model nested within morph). One data set incorporated only the SNPs with significantly different allele frequencies between morphs, while the other included subsets of randomly selected SNPs not significantly differentiated between morphs. For the latter data set, three random subsets of 10 000 SNPs were generated through the procedure reported in the Stacks documentation after removing those SNPs whose allele frequencies significantly differentiated between morphs. We did not exclude the SNPs significantly differentiated in both the comparison between morphs and between sites from these two types of AMOVA data sets. In these AMOVA analyses, we applied a hierarchical

study design in which locations were nested within morphs. Following this scheme, genetic variation is partitioned in three components: among morphs, among locations within morphs and among individuals within locations.

PoolSeq bioinformatic pipelines

SEQPREP v. 1.1 (https://github.com/jstjohn/SeqPrep) and CLC Genomics Workbench v. 8.0.2 (CLC bio, Aarhus, Denmark) were used to remove adapters and trim raw PoolSeq Illumina reads at 151 bp. These were mapped individually for each pool to the existing cichlid fish (*O. niloticus*, *M. zebra*, *Pundamilia nyererei*, *Neolamprologus brichardi* and *Astatotilapia burtoni*) reference genomes (Brawand *et al.* 2014) using BWA v. 0.7.12 (Li & Durbin 2009) and BOWTIE2 v. 2.2.5 (Langmead & Salzberg 2012) using both default and optimized settings. These include maximum edited distance (-*n*) 0.01, seed (-*l*) 100, maximum number of gap opens (-*l*) 2, disallow long deletion within 12 bp towards 3′ end (-*d*) and maximum number of gap extensions (-*e*) 12. Mapped pools belonging to the same morph or site were merged through CLC Genomics Workbench, obtaining two (left and right; morph data set) or four (Katoto, Kasakalawe, Mbita and Toby; geographic data set) pools for subsequent analyses. SAMTOOLS v. 1.2 (Li *et al.* 2009) and PICARD v. 1.119 (http://picard.sourceforge.net.) were used to remove duplicates and low-quality alignments (mapping quality lower than 20; unmapped reads or without both mates aligned to the reference genome). The resulting files were exported to a single mpileup file containing the pools to be compared without quality score adjustment. Indels and repetitive regions were masked considering a window of five nucleotides through POPOOLATION v. 1.2.2 (Kofler *et al.* 2011a). A sync-file was built using POPOOLATION2 v. 1.2.01 (Kofler *et al.* 2011b), with a minimum base quality of 20, followed by subsampling without replacement to a target coverage of 10, minor allele count of 2 and maximum coverage of 200. POPOOLATION2 was also used to calculate the fixation index F_{ST} (Hartl & Clark 2007) and to test for differences in allele frequencies using the Fisher's exact test (Fisher 1922). Together with single-SNP analyses, we also performed analyses using nonoverlapping sliding windows of 100 bp, a minimum count of 3 and a minimum covered fraction of 1 (i.e. the entire 100-bp sequence of a given window had to be present) to minimize stochastic errors (Kofler *et al.* 2011a). Corrections for multiple tests and exclusion of SNPs differentiated in both the morph and the geographic data set were performed as described in the ddRAD data set. PCA was performed in R using the overall F_{ST} values between locations.

Gene prediction and functional annotation

To annotate the regions significantly associated with mouth asymmetry, the following procedure was applied: (i) for the ddRAD data set, the consensus sequence of the locus containing the significant SNP was aligned to the *M. zebra* genome using BLASTN v. 2.2.30 (Altschul *et al.* 1997) with an e-value threshold of 1e-35. Given the relative short size of the scaffold to which the RAD tag aligned to (scaffold 554; 50 966 bp), all the genes included here were retrieved from the available annotation. (ii) For the PoolSeq data set, as the *M. zebra* genome was used as reference in the PoPoolation analysis, this mapping information was implemented to retrieve the genes (again using the *M. zebra* annotation) included upstream and downstream (±10 000 bp) the location of the significant SNPs (sliding windows). For both data sets, the genes were further functionally annotated using BLASTX and BLAST2GO v. 2.8 (Conesa *et al.* 2005) using default settings and the lowest Gene Ontology level. The presence of significant GO term frequency differences in the genes occurring in the identified regions was tested comparing the PoolSeq gene sets with a baseline including all the *O. niloticus* genes. For this purpose, the BLAST2GO enrichment analysis was implemented using the Fisher's exact test and setting the false discovery rate to 0.05 (Benjamini & Yekutieli 2001).

Results

ddRAD

Illumina sequencing generated 128 820 739 raw reads. After filtering, we retained 109 387 016 reads. The *de novo* pipeline identified 155 798 SNPs, reduced to 76 836 after the *rxstacks* correction and filtering for coverage.

After correcting for multiple tests, only a single SNP was significantly differentiated between the left and right morph fish (F_{ST} 0.8134; BH and SB corrected *P*-value 0.000154; Fig. 2). This SNP was excluded from the geographic comparison as it deviated from HWE. The same SNP was retrieved in the *de novo* assemblies performed excluding Congo specimens or using different parameters (data not shown), except the data sets having -n (*cstacks*) set to 0, -m and -M (*ustacks*) higher than five and three, that did not produce significant SNPs after multiple test correction. This SNP presented two alternative nucleotides: G, predominant in the right group, and A mostly related to the left morph (Table S2, Supporting information). The ddRAD locus containing this SNP aligned to the *Maylandia zebra* (unplaced genomic scaffold 554; 50 966 bp; score 262;

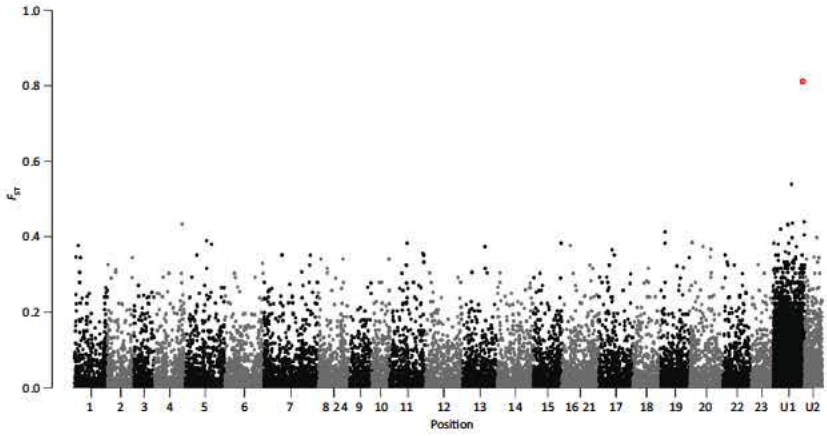

Fig. 2 Manhattan plot of F_{ST} between morphs in the ddRAD data set. The SNP significant after correcting for multiple tests is highlighted in red (empty circle). Numbers 1 23 refer to the corresponding linkage groups in the *Oreochromis niloticus* genome; U1 refers to unplaced scaffolds; U2 to SNPs in sequences that did not blast neither on the *O. niloticus* nor on the *Maylandia zebra* genomes. These were, then, randomly ordered.

similarity percentage 96%; E-value 2e-58) and on the *Pundamilia nyererei* (unplaced genomic scaffold 3817; 2740 bp; score 252; similarity percentage 95%; E-value 1e-55) genomes. The *P. nyererei* scaffold falls within the *M. zebra* one, coinciding with the same genomic region (score 5100; similarity percentage 97%; *E*-value 0.0), which includes three genes and one pseudogene related to immunity response, specifically the immunoglobulin light chain (Table S3, Supporting information).

A mean of 40 245 (standard deviation 5670) SNPs after removal of loci significantly deviating from HWE were analysed to assess genetic variation in geographic space. Pairwise comparisons between the geographic sites resulted to be all significant after multiple test correction (Table 1). The overall F_{ST} value increased with increasing geographic distance (Table 1). The PCA result (Fig. 3a) suggested that most of genetic variation is found between the sampling sites in Congo and the rest. There is also a certain level of variation among the four Zambia sites but with a considerable overlap between Kasakalawe and Mbita.

The AMOVA analysis using only the SNP with significant difference in allele frequencies between morphs indicated that the among-morphs term was significant and accounted for 16.29% of variation. On the other hand, differentiation between locations within morphs was lower and not significant (Table S4, Supporting information). On the contrary, the random subsets did not show significant structuring between morphs but among locations within morphs (Table S4, Supporting information). The among-individuals within-locations source of variation was significant in all data sets.

Table 1 Pairwise F_{ST} between sampling locations. In the upper triangle are reported the values obtained with the PoolSeq data set while the F_{ST} obtained with the ddRAD data set are in the lower triangle. Congo was excluded from the PoolSeq data set. All the comparisons were significant after correcting for multiple tests

PoolSeqddRAD	Katoto	Kasakalawe	Mbita	Toby	Congo
Katoto		0.0206	0.0223	0.0267	
Kasakalawe	0.0312		0.0184	0.0219	
Mbita	0.0289	0.0106		0.0223	
Toby	0.0723	0.0472	0.0437		
Congo	0.2314	0.2280	0.2147	0.2819	

PoolSeq

We obtained between 18 613 620 and 26 095 562 (mean 22 371 737; standard deviation 3 431 353) raw reads per pool from Illumina sequencing. Remarkably, we obtained a similar number of raw reads between the eight pools, essential to analyse them effectively (Schlötterer *et al.* 2014). Trimming and cleaning resulted in between 18 500 590 and 26 066 400 (mean 22 323 225; standard deviation 3 447 229) reads per pool. No appreciable improvement was observed between mapping using the default and optimized parameters (data not shown); subsequently, the default settings were used for the following steps. Mean alignment rates across pools were 80.36% (*M. zebra*; standard deviation 0.65), 68.20% (*Oreochromis niloticus*; standard deviation 0.57), 78.94% (*P. nyererey*; standard deviation 0.65), 75.63% (*Neolamprologus brichardi*; standard deviation 0.62) and 79.68% (*Astatotilapia burtoni*; standard deviation 0.62). Consequently, the *M. zebra* assembly was used for subsequent analyses.

We identified 3 970 889 SNPs. These were reduced to 755 810 (single-SNP analysis) and 61 270 (100-bp sliding window approach) after filtering for quality and coverage. After correcting for multiple tests, the single-SNP analysis did not produce any significant SNP in the comparison between morphs, as well as in the pairwise comparison between geographic locations. Interestingly, the 100-bp data set resulted in 395 (after the BH multiple test correction procedure) and 38 (applying the SB method) windows containing SNPs significantly

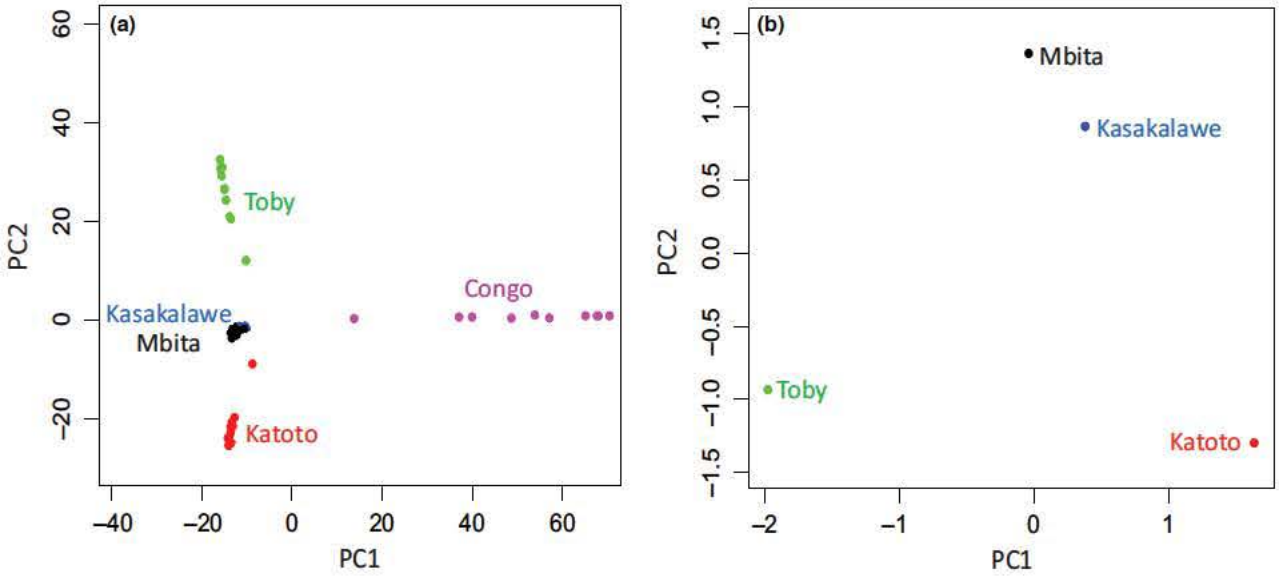

Fig. 3 Plot of the scores along the first two principal components of the ddRAD (a) and PoolSeq (b) data sets. Congo was excluded from the PoolSeq data set.

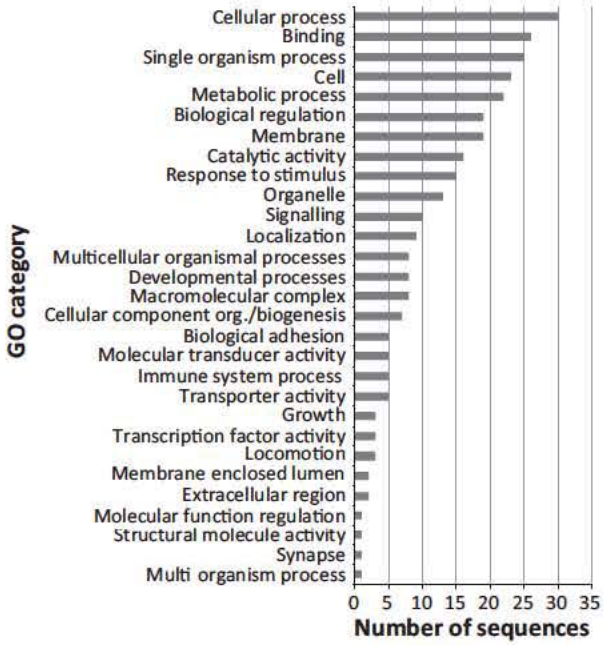

Fig. 4 Summary of the GO terms for the PoolSeq sequences containing the significant SNPs after correcting for multiple tests with the Benjamini Hochberg procedure and removing the SNPs significantly differentiated when comparing sampling sites.

differentiated between the left and right samples. Seventeen of 395 windows of the BH data set included SNPs whose frequencies were significantly different among locations. For this reason, these windows were excluded from subsequent analyses. The functional annotation of the resulting 378 loci identified 108 (BH) and 22 (SB) genes with known function (Figs 4 and 5; Tables S5 and S6, Supporting information). These genes were significantly enriched for several functions when the Nile tilapia (*O. niloticus*) genome was used as background (Figs S1 and S2, Supporting information), particularly representatives related to response to stimuli, immunity (BH), cell adhesion and *trans*membrane signalling pathway (BH and SB).

The geographic comparison showed, as expected, higher differentiation at larger geographic distance (Table 1; Fig. 3b).

Discussion

Perissodus microlepis is an outstanding example of morphological and behavioural laterality and a textbook model of negative frequency-dependent selection. However, the processes producing and maintaining this left right asymmetry remain unclear. Our results suggest that the notable polymorphism in *P. microlepis* has a significant genetic basis, in particular a polygenic contribution, and that geographic structure needs to be taken into consideration in the attempt to identify genetic loci differentiated between morphs.

Molecular markers and mapping

This study represents the first genome-wide analysis of *P. microlepis* intraspecific genetic diversity. Previous studies of the genetic variation in this cichlid had used

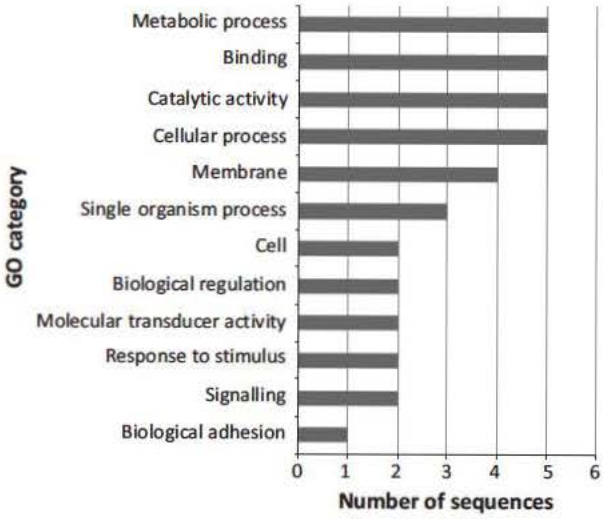

Fig. 5 Summary of the GO terms for the PoolSeq sequences containing the significant SNPs after correcting for multiple tests with the sequential Bonferroni procedure and removing the SNPs significantly differentiated when comparing sampling sites.

the mitochondrial control region (Koblmüller *et al.* 2009; Lee *et al.* 2010), or relatively few (13 in Lee *et al.* 2010; five in Stewart & Albertson 2010) microsatellite loci. These previous analyses, involving few genomic regions, were therefore limited in power by the number and type of the chosen markers, and asked different questions. Thanks to the rapid development and decreasing costs of high-throughput DNA sequencing technologies in the last years, we were able to obtain more than 150 000 (ddRADseq) and 3 900 000 (PoolSeq) SNPs. Additionally, the combination of individual and pooled sequencing enabled us to obtain a larger number of markers throughout the genome than any single technique would have.

The accuracy of mapping of the PoolSeq data set on different genomes reflects the time of divergence between *P. microlepis* and each of the five African cichlid species with published reference genomes (Brawand *et al.* 2014). *Neolamprologus brichardi* is the only cichlid endemic to Lake Tanganyika among the five with reference genomes; however, it is not the most closely related species to the *Perissodini* lineage. Rather, among the African cichlids lineages with published genomes, haplochromine cichlids (such as *M. zebra* or *A. burtoni*) are more closely related to *P. microlepis* than some of the other tribes of cichlids that are endemic to Lake Tanganyika (Salzburger *et al.* 2005; Brawand *et al.* 2014). Perhaps not surprisingly, *Oreochromis niloticus*, a cichlid that has the best genome sequence published so far, but is phylogenetically distant to *P. microlepis*, had the worst mapping accuracy.

Genetic bases of mouth asymmetry

The *de novo* ddRAD assemblies using several parameters were all concordant in the identification of one SNP significantly differentiated between the left and right groups. Three assembly parameter settings distance allowed between catalogue loci (*cstacks* -n) 0, minimum stack depth (*ustacks* -m) greater than 5 and distance allowed between stacks (*ustacks* -M) higher than 3 did not identify any significant SNP. However, assemblies using these three settings are likely not appropriate, as reported in the Stacks manual. Considering that the results of the analyses using the remaining wide range of parameters and combinations are concordant with each other, we were confident that the SNP we found significantly differentiated between morphs did not result from inappropriate assembly settings but represents a true polymorphism. Unfortunately, it was not possible to evaluate this SNP through PoolSeq as this locus was discarded during the filtering procedure due to low coverage.

We did not detect any SNP that was significantly differentiated between morphs in the single-SNP analysis of the PoolSeq data set. However, significant SNPs were also absent in the PoolSeq analysis of geographic variation. This clearly contrasts with the results from this (ddRAD and 100-bp PoolSeq data sets) and previous (Koblmüller *et al.* 2009; Lee *et al.* 2010) studies, which agree in reporting significant genetic divergence across geographic locations. Additionally, the 100-bp data set, implementing more restrictive filtering parameters and thus resulting in a lower number of higher-quality SNPs, produced 378 (BH) and 38 (SB) windows containing SNPs that were significantly differentiated between morphs. These findings suggest that the absence of significant SNPs in the single-SNP analysis is more likely to be a consequence of the applied procedures and does not reflect the real pattern of differentiation between morphs. In the ddRAD data set, we found only one significant SNP. Probably this is related to the notable restrictiveness of the multiple test correction, and it is likely that there are more SNPs underlying the left right polymorphism. In fact, while it is recommended to control for the type I error rate, many of these methods are rather conservative (Shaffer 1995; Ge *et al.* 2003; Moran 2003; Camargo *et al.* 2008; Carvajal-Rodríguez *et al.* 2009; Benjamini 2010). Alternatively, the discrepancy between the number of significant SNPs obtained with the 1-bp or 100-bp windows approaches might suggest that multiple SNPs affect the gene(s) underlying the trait, but each SNP alone does not contribute enough to be detected. Finally, the different numbers of significant SNPs in the PoolSeq 100-bp BH and SB analyses are due to the different level of conservativeness of

the BH and SB method. BH seemed to be better suited to our study having a high number of tests, but SB provides more stringent results, although it is prone to strongly underestimate the number of SNPs truly differentiated between the left and right morph.

Interestingly, both ddRAD and PoolSeq marker data sets analysed here indicated the presence of genes related to immunity in the genomic regions differentiated between morphs. Immunoglobulin (ddRAD) and major histocompatibility complex (MHC; PoolSeq) have already been proposed as a potent factor contributing to the divergence of cichlids lineages, and promising candidates for the analysis of functional relevance with regard to phenotype and divergence (Machado *et al.* 2014 and references therein). MHC is known to contribute to both assortative and disassortative mating in closely related cichlids and other fishes (e.g. Landry *et al.* 2001; Reusch *et al.* 2001), and consequently, these genes have been suggested as one of the mechanisms of adaptive ecological speciation (Piertney & Oliver 2006; Blais *et al.* 2007; Salzburger 2009; Eizaguirre & Lenz 2010; Eizaguirre *et al.* 2011; Evans *et al.* 2012 and reference therein). Our result suggests that these might contribute also to nonrandom mating between the left and right morph of *P. microlepis*. To date, contradictory findings exist on the presence of assortative, disassortative or random mating in *P. microlepis* (Takeuchi & Hori 2008; Lee *et al.* 2010; Kusche *et al.* 2012). On one hand, disassortative mating has been advocated to have a role in stabilizing the mouth polymorphism (Takeuchi & Hori 2008), while other studies did not detect any signature of selective mating and concluded that random mating occurs in natural populations of *P. microlepis* (Lee *et al.* 2010; Kusche *et al.* 2012). However, nonrandom mating is not expected to have a genome-wide effect, but should only affect loci involved in selective mating choice, and regions closely linked to them (Templeton 2006). This might explain the absence of any obvious genetic signature of nonrandom mating in a data set based on a small number of markers (mitochondrial control region and 13 microsatellites; Lee *et al.* 2010) compared to our work. It is possible that among the genes that were identified as potential candidate genes underlying mouth laterality (Tables S5 and S6, Supporting information) there are genes involved in nonrandom mating.

Perhaps more interestingly, the analyses of the PoolSeq data set were concordant in finding genes involved in cell adhesion, particularly the protocadherins, in the regions with different allele frequencies between morphs. Protocadherins are a subgroup of the cadherin superfamily of homophilic cell adhesion proteins (Hulpiau & Van Roy 2009 and references therein). Adhesion molecules regulate cellular migration and allow the direct transfer of small molecule signals. Cellular movement and communication is at the basis of the mechanisms determining the early establishment of the left right patterning during embryogenesis (Burdine & Schier 2000; Mercola & Levin 2001; Levin 2005 and references therein). Additionally, PoolSeq BH results indicated the presence of several genes related to ion transporter activity. The chief role of both transporter and adhesion molecules in the left right development has been demonstrated in gain- and loss-of-function experiments, in which expression alterations of these proteins randomize the left right axis (Levin 2005 and references therein). In fact, the initial break of symmetry is caused by an asymmetrical transmission of the positional information (in form of signalling molecules or ion flux; Levin 2005). This results in the accumulation of a determinant on one side of the developing embryo (e.g. *Shh* on the chicken left side; Burdine & Schier 2000), which, in turn, determines the cascade of asymmetric gene expression leading to the differentiation of the left and right margins (Levin 2005 and references therein). Cadherins are one of the earliest proteins to be asymmetrically expressed in the chick embryo and have been suggested to specify cell polarity (Garćia-Castro *et al.* 2000; Levin 2005 and references therein). Protocadherins are predominantly expressed in the brain and are involved in neural network formation (Sano *et al.* 1993). In humans, the origin of cerebral asymmetry and language has been related to these genes, and their mutations have been associated with schizophrenia and neurodegenerative illness (Anderton *et al.* 1998; Kalmady & Venkatasubramanian 2009 and references therein). In fish, cerebral asymmetry is linked to handed behaviour (e.g. Reddon *et al.* 2009; Takeuchi *et al.* 2010; Concha *et al.* 2012 and references therein). Lateralized feeding behaviour is probably expressed earlier in development than mouth asymmetry in *P. microlepis*, as two-month-old fishes already exhibit handed behaviour and attack-side preference (Lee *et al.* 2012). It has been proposed that lateralized behaviour precedes and facilitates mouth asymmetry (Van Dooren *et al.* 2010; Lee *et al.* 2012) and that the genetic basis of this trait would primarily affect behavioural laterality rather than morphology (Van Dooren *et al.* 2010; Lee *et al.* 2012). Our results support this hypothesis, suggesting that protocadherins might play a central role in the establishment of *P. microlepis* asymmetry *via* behavioural lateralization due to their key function in cerebral asymmetry. Alternatively, the regions containing the significant SNPs might not harbour the causal genes of mouth asymmetry, but only be genetically linked to them.

Taken together, our results suggest a sizable and polygenic basis of mouth asymmetry. This is in agreement with previous studies proposing that this trait is

unlikely to be determined by a single genetic locus with two alleles and does not follow simple Mendelian inheritance (Kusche *et al.* 2012; Lee *et al.* 2015).

Geographic structuring

A significant genetic variation was observed among all the sampling sites, even at small spatial scale. This is in agreement with previous phylogeographic studies (Koblmüller *et al.* 2009; Lee *et al.* 2010).

The presence of population stratification is one of the well-known sources of false positives in studies associating phenotypic and genotypic information. Several methods have been proposed to deal with this problem in association mapping: genomic control, principal component analysis, structured association analysis and mixed models. Each of them has critical limitations, such as the high rate of false negatives (Ehrenreich *et al.* 2009; Shin & Lee 2015; Wellenreuther & Hansonn 2016 and references therein). Here, we used a simple but effective procedure to control for geographic structuring: we controlled for geographic provenance in a statistical model, and let the results of the analysis of variation in geographic space inform the analysis of variation between morphs. AMOVA (where genetic variation is decomposed in terms, in this case corresponding to variation between morphs and variation between sampling sites) confirmed that the SNP significantly different between morphs in the ddRAD data set is not a false positive due to geographic structuring. Similarly, the PoolSeq SB resulted to be free of spurious genetic association due to geographic stratification. On the other hand, the PoolSeq BH candidate SNPs included 17 windows holding SNPs significant also in the comparisons between sampling sites. Analysing differentiation between morphs disregarding genetic variation across the geographic space would have probably resulted in the inclusion of false positives. On the contrary, we discarded the SNPs whose frequencies were significantly different both between morphs and between sampling sites, thus reducing the chance of false positives. These findings also highlight the importance of considering the influence of geographic stratification together with other sources of spurious associations if known in studies with designs and goals similar to ours as such analyses are increasingly feasible due to the reduction in costs of genomewide sequencing technologies.

An alternative approach to prevent the influence of geographic structuring involves comparing the left and right morph within each sampling location. This would also allow testing the fascinating hypothesis of differences in genetic determination between sites due to developmental system drift (i.e. development of homologous traits *via* divergent mechanisms; True & Haag 2001), a *scenario* which has not been previously considered. Indeed, to date all the studies on *P. microlepis*, including this one, assumed a common genetic basis for mouth asymmetry across populations. This assumption constitutes, then, a null hypothesis that should be properly tested in future studies based on larger intrapopulation samples.

Conclusions

This study provides the first insight into the genomic architecture of *Perissodus microlepis* mouth asymmetry. Importantly, it clarified that this interesting trait has a genetic basis, which is likely to be influenced by multiple loci. The presence of many differentiated loci between the most right and most left individuals in natural populations contradicts both the hypothesis of no genetic determination and the single locus genetic model, but confirms recent findings suggesting a quantitative architecture of mouth asymmetry. Further, we describe a set of candidate genomic regions while controlling for false positives due to geographic stratification. While we are far from a complete understanding of the genotype phenotype map of this iconic trait, our data provide an important contribution to a deeper understanding of left right asymmetry and the processes driving the evolution and maintenance of intraspecific polymorphisms in animals.

Acknowledgements

We thank Henrik Kusche for help with collecting fishes in 2010, Lènia da Conceicao Ferrao Beck for laboratory assistance, Andreas Kautt for assistance with the ddRAD analysis and other Meyer laboratory members for their helpful suggestions. We are grateful to three anonymous reviewers for their valu able comments and suggestions on the first version of the manuscript. FR is funded by the International Max Planck Research School (IMPRS) for Organismal Biology and the DAAD (scholarship 2015/16 57130104). CF was funded by a Marie Curie IEF Fellowship (Grant Agreement 327875 Plastic itySpeciation). PF is financially supported by a German Research Foundation (DFG) Research Grant (DFG15957314). The University of Konstanz is thanked for its support to the Meyer laboratory and the GeCKo (Genomic Center Konstanz). Funding for this project came from DFG grant ME1725 18 (to Hyuk Je Lee and AM).

References

Altschul SF, Madden TL, Schaffer AA *et al.* (1997) Gapped BLAST and PSI BLAST: a new generation of protein data base search programs. *Nucleic Acids Research*, **25**, 3389 3402.

Anderton BH, Callahan L, Coleman P *et al.* (1998) Dendritic changes in Alzheimer's disease and factors that may under lie these changes. *Progress in Neurobiology*, **55**, 595 609.

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

Balding DJ (2006) A tutorial on statistical methods for popula tion association studies. *Nature Reviews Genetics*, **7**, 781 791.

Benjamini Y (2010) Discovering the false discovery rate. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **72**, 405 416.

Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B (Methodological)*, **57**, 289 300.

Benjamini Y, Yekutieli D (2001) The control of the false discov ery rate in multiple testing under dependency. *Annals of Sta tistics* **29**, 1165 1188.

Blais J, Rico C, van Oosterhout C *et al.* (2007) MHC adaptive divergence between closely related and sympatric African Cichlids. *PLoS ONE*, **2**, e734.

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

Burdine RD, Schier AF (2000) Conserved and divergent mecha nisms in left right axis formation. *Genes & Development*, **14**, 763 776.

Camargo A, Azuaje F, Wang H, Zheng H (2008) Permutation based statistical tests for multiple hypotheses. *Source Code for Biology and Medicine*, **3**, 15.

Carvajal Rodriguez A, de Una Alvarez J (2011) Assessing sig nificance in high throughput experiments by sequential goodness of fit and q value estimation. *PLoS ONE*, **6**, e24700.

Carvajal Rodríguez A, de Una Alvarez J, Rolán Alvarez E (2009) A new multitest correction (SGoF) that increases its statistical power when increasing the number of tests. *BMC Bioinformatics*, **10**, 209.

Catchen J, Hohenlohe PA, Bassham S, Amores A, Cresko WA (2013) Stacks: an analysis tool set for population genomics. *Molecular Ecology*, **22**, 3124 3140.

Concha ML, Bianco IH, Wilson SW (2012) Encoding asymme try within neural circuits. *Nature Reviews Neuroscience*, **13**, 832 843.

Conesa A, Gotz S, García Gómez JM *et al.* (2005) Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics*, **21**, 3674 3676.

Cook LM, Saccheri IJ (2013) The peppered moth and industrial melanism: evolution of a natural selection case study. *Hered ity*, **110**, 207 212.

Darwin C (1859) *On the Origins of Species by Means of Natural Selection*. Murray, London.

Ehrenreich I, Gerke J, Kruglyak L (2009) Genetic dissection of complex traits in yeast: insights from studies of gene expres sion and other phenotypes in the BY× RM cross. *Cold Spring Harbor Symposia on Quantitative Biology*, **74**, 145 153.

Eizaguirre C, Lenz T (2010) Major histocompatibility complex polymorphism: dynamics and consequences of parasite mediated local adaptation in fishes. *Journal of Fish Biology*, **77**, 2023 2047.

Eizaguirre C, Lenz TL, Sommerfeld RD *et al.* (2011) Parasite diversity, patterns of MHC II variation and olfactory based mate choice in diverging three spined stickleback ecotypes. *Evolutionary Ecology*, **25**, 605 622.

Evans ML, Dionne M, Miller KM, Bernatchez L (2012) Mate choice for major histocompatibility complex genetic divergence as a bet hedging strategy in the Atlantic salmon (*Salmo salar*). *Proceedings: Biological Sciences*, **279**, 379 386.

Excoffier L, Lischer HE (2010) Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Molecular Ecology Resources*, **10**, 564 567.

Excoffier L, Smouse PE, Quattro JM (1992) Analysis of molecu lar variance inferred from metric distances among DNA hap lotypes application to human mitochondrial DNA restriction data. *Genetics*, **131**, 479 491.

Fisher RA (1922) On the interpretation of $\chi 2$ from contingency tables, and the calculation of P. *Journal of the Royal Statistical Society*, **85**, 87 94.

Fisher SRA (1958) *The Generical Theory of Natural Selection*, 2nd edn. Dover, New York.

Fleiss J, Shrout P (1977) The effects of measurement errors on some multivariate procedures. *American Journal of Public Health*, **67**, 1188 1191.

Franchini P, Fruciano C, Spreitzer ML *et al.* (2014) Genomic architecture of ecologically divergent body shape in a pair of sympatric crater lake cichlid fishes. *Molecular Ecology*, **23**, 1828 1845.

Freedman ML, Reich D, Penney KL *et al.* (2004) Assessing the impact of population stratification on genetic association studies. *Nature Genetics*, **36**, 388 393.

Fruciano C (2016) Measurement error in geometric morpho metrics. *Development Genes and Evolution*, 1 20. doi: 10.1007/ s00427 016 0537 4.

Fruciano C, Tigano C, Ferrito V (2011a) Geographical and mor phological variation within and between colour phases in *Coris julis* (L. 1758), a protogynous marine fish. *Biological Journal of the Linnean Society*, **104**, 148 162.

Fruciano C, Tigano C, Ferrito V (2011b) Traditional and geo metric morphometrics detect morphological variation of lower pharyngeal jaw in *Coris julis* (Teleostei, Labridae). *Ital ian Journal of Zoology*, **78**, 320 327.

Fruciano C, Tigano C, Ferrito V (2012) Body shape variation and colour change during growth in a protogynous fish. *Environmental Biology of Fishes*, **94**, 615 622.

Futschik A, Schlotterer C (2010) The next generation of molecu lar markers from massively parallel sequencing of pooled DNA samples. *Genetics*, **186**, 207 218.

Futuyma DJ (2009) *Evolution*, 2nd edn. Sinauer Associates INC, Sunderland, Massachusetts.

Garćia Castro MNI, Vielmetter E, Bronner Fraser M (2000) N Cadherin, a cell adhesion molecule involved in establishment of embryonic left right asymmetry. *Science*, **288**, 1047 1051.

Ge Y, Dudoit S, Speed TP (2003) Resampling based multiple testing for microarray data analysis. *Test*, **12**, 1 77.

Hartl DL, Clark A (2007) *Principles of Population Genetics*. Sin auer, Sunderland, Massachusetts.

Hata H, Hori M (2012) Inheritance patterns of morphological laterality in mouth opening of zebrafish, *Danio rerio*. *Lateral ity: Asymmetries of Body Brain and Cognition*, **17**, 741 754.

Hata H, Takahashi R, Ashiwa H *et al.* (2012) Inheritance patterns of lateral dimorphism examined through breeding experiments in tanganyikan cichlid (*Julidochromis transcriptus*) and Japanese Medaka (*Oryzias latipes*). *Zoological Science*, **29**, 49 53.

Henning F, Meyer A (2014) The evolutionary genomics of cich lid fishes: explosive speciation and adaptation in the postge nomic era. *Annual Review of Genomics and Human Genetics*, **15**, 417 441.

Holm S (1979) A simple sequentially rejective multiple test pro cedure. *Scandinavian Journal of Statistics*, **6**, 65 70.

Hori M (1993) Frequency dependent natural selection in the handedness of scale eating cichlid fish. *Science*, **260**, 216 219.

Hori M (2000) Gunshuu no tayousei to anteika kikou (The diversities and stabilizing mechanisms of communities). In: *Gunshuu Seitaigaku no Genzai (Current Community Ecology)* (eds Sato H, Yamamoto T), pp. 257 283. Kyoto University Press, Kyoto.

Hori M, Ochi H, Kohda M (2007) Inheritance pattern of lateral dimorphism in two cichlids (a scale eater, *Perissodus microle pis*, and an herbivore, *Neolamprologus moorii*) in Lake Tan ganyika. *Zoological Science*, **24**, 486 492.

Hubbs C, Hubbs L (1945) Bilateral asymmetry and bilateral variation in fishes. *Papers from the Michigan Academy of Science, Arts and Letters*, **30**, 229 311.

Hulpiau P, Van Roy F (2009) Molecular evolution of the cad herin superfamily. *The international Journal of Biochemistry & Cell Biology*, **41**, 349 369.

Jombart T, Ahmed I (2011) adegenet 1.3 1: new tools for the analysis of genome wide SNP data. *Bioinformatics*, **27**, 3070 3071.

Kalmady SV, Venkatasubramanian G (2009) Evidence for posi tive selection on protocadherin Y gene in *Homo sapiens*: implications for schizophrenia. *Schizophrenia Research*, **108**, 299 300.

Koblmuller S, Egger B, Sturmbauer C, Sefc KM (2007) Evolu tionary history of Lake Tanganyika's scale eating cichlid fishes. *Molecular Phylogenetics and Evolution*, **44**, 1295 1305.

Koblmuller S, Duftner N, Sefc KM *et al.* (2009) Phylogeographic structure and gene flow in the scale eating cichlid *Perissodus microlepis* (Teleostei, Perciformes, Cichlidae) in southern Lake Tanganyika. *Zoologica Scripta*, **38**, 257 268.

Kofler R, Orozco terWengel P, De Maio N *et al.* (2011a) PoPoolation: a toolbox for population genetic analysis of next generation sequencing data from pooled individuals. *PLoS ONE*, **6**, e15925.

Kofler R, Pandey RV, Schlotterer C (2011b) PoPoolation2: iden tifying differentiation between populations using sequencing of pooled DNA samples (Pool Seq). *Bioinformatics*, **27**, 3435 3436.

Kusche H, Lee HJ, Meyer A (2012) Mouth asymmetry in the textbook example of scale eating cichlid fish is not a discrete dimorphism after all. *Proceedings of the Royal Society B Biologi cal Sciences*, **279**, 4715 4723.

Landry C, Garant D, Duchesne P, Bernatchez L (2001) 'Good genes as heterozygosity': the major histocompatibility com plex and mate choice in Atlantic salmon (*Salmo salar*). *Pro ceedings of the Royal Society of London B: Biological Sciences*, **268**, 1279 1285.

Langmead B, Salzberg SL (2012) Fast gapped read alignment with Bowtie 2. *Nature Methods*, **9**, 357 359.

Lee HJ, Pittlik S, Jones JC *et al.* (2010) Genetic support for random mating between left and right mouth morphs in the dimorphic scale eating cichlid fish *Perissodus microlepis* from Lake Tanganyika. *Journal of Fish Biology*, **76**, 1940 1957.

Lee HJ, Kusche H, Meyer A (2012) Handed foraging behavior in scale eating cichlid fish: its potential role in shaping mor phological asymmetry. *PLoS ONE*, **7**, 8.

Lee HJ, Heim V, Meyer A (2015) Genetic and environmental effects on the morphological asymmetry in the scale eating cichlid fish, *Perissodus microlepis*. *Ecology and Evolution*, **5**, 4277 4286.

Levin M (2005) Left right asymmetry in embryonic develop ment: a comprehensive review. *Mechanisms of Development*, **122**, 3 25.

Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows Wheeler transform. *Bioinformatics*, **25**, 1754 1760.

Li H, Handsaker B, Wysoker A *et al.* (2009) The sequence alignment/map format and SAMtools. *Bioinformatics*, **25**, 2078 2079.

Machado HE, Jui G, Joyce DA *et al.* (2014) Gene duplication in an African cichlid adaptive radiation. *BMC Genomics*, **15**, 161.

Magwene PM, Willis JH, Kelly JK (2011) The statistics of bulk segregant analysis using next generation sequencing. *PLoS Computational Biology*, **7**, e1002255.

Mboko S, Kohda M, Hori M (1998) Asymmetry of mouth open ing of a small herbivorous cichlid fish *Telmatochromis tempo ralis* in Lake Tanganyika. *Zoological Science*, **15**, 405 408.

Mercola M, Levin M (2001) Left right asymmetry determina tion in vertebrates. *Annual Review of Cell and Developmental Biology*, **17**, 779 805.

Meyer A (2015) Extreme evolution. *Scientific American*, **312**, 70 75.

Michelmore RW, Paran I, Kesseli R (1991) Identification of mark ers linked to disease resistance genes by bulked segregant analysis: a rapid method to detect markers in specific genomic regions by using segregating populations. *Proceedings of the National Academy of Sciences USA*, **88**, 9828 9832.

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.

Moran MD (2003) Arguments for rejecting the sequential bon ferroni in ecological studies. *Oikos*, **100**, 403 405.

Nakajima M, Matsuda H, Hori M (2004) Persistence and fluctu ation of lateral dimorphism in fishes. *American Naturalist*, **163**, 692 698.

Nakajima M, Yodo T, Katano O (2007) Righty fish are hooked on the right side of their mouths observations from an angling experiment with largemouth bass, *Micropterus sal moides*. *Zoological Science*, **24**, 855 859.

Nielsen R, Beaumont MA (2009) Statistical inferences in phylo geography. *Molecular Ecology*, **18**, 1034 1047.

Nshombo M, Yanagisawa Y, Nagoshi M (1985) Scale eating in *Perissodus microlepis* (Cichlidae) and change of its food habits with growth. *Japanese Journal of Ichthyology*, **32**, 66 73.

Palmer AR (2004) Symmetry breaking and the evolution of development. *Science*, **306**, 828 833.

Palmer AR (2005) Chapter 16 Antisymmetry A2 Hallgríms son, Benedikt. In: *Variation* (ed. Hall BK), pp. 359 XXIV. Aca demic Press, Burlington.

Palmer AR (2010) Scale eating cichlids: from hand (ed) to mouth. *Journal of Biology*, **9**, 11.

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.

Piertney S, Oliver M (2006) The evolutionary ecology of the major histocompatibility complex. *Heredity*, **96**, 7 21.

Price AL, Patterson NJ, Plenge RM *et al.* (2006) Principal com ponents analysis corrects for stratification in genome wide association studies. *Nature Genetics*, **38**, 904 909.
Pritchard JK, Donnelly P (2001) Case control studies of associa tion in structured or admixed populations. *Theoretical Popula tion Biology*, **60**, 227 237.
Purcell S, Neale B, Todd Brown K *et al.* (2007) PLINK: a tool set for whole genome association and population based linkage analyses. *The American Journal of Human Genetics*, **81**, 559 575.
R core team (2013) *A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
Reddon AR, Gutiérrez Ibánez C, Wylie DR, Hurd PL (2009) The relationship between growth, brain asymmetry and behavioural lateralization in a cichlid fish. *Behavioural Brain Research*, **201**, 223 228.
Reusch TB, HaÈberli MA, Aeschlimann PB, Milinski M (2001) Female sticklebacks count alleles in a strategy of sexual selec tion explaining MHC polymorphism. *Nature*, **414**, 300 302.
Robinson B, Schluter D (2000) Natural selection and the evolu tion of adaptive genetic variation in northern freshwater fishes. In: *Adaptive Genetic Variation in the Wild*, (eds Mous seau TA, Sinervo B, Endler J), pp. 65 94. Oxford University Press, Oxford.
Rohlf F (2006) *TpsDig, Version 2.10*. Department of Ecology and Evolution, State University of New York, Stony Brook.
Salzburger W (2009) The interaction of sexually and naturally selected traits in the adaptive radiations of cichlid fishes. *Molecular Ecology*, **18**, 169 185.
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.
Sano K, Tanihara H, Heimark RL *et al.* (1993) Protocadherins: a large family of cadherin related molecules in central nervous system. *The EMBO Journal*, **12**, 2249.
Schlotterer C, Tobler R, Kofler R, Nolte V (2014) Sequencing pools of individuals mining genome wide polymorphism data without big funding. *Nature Reviews Genetics*, **15**, 749 763.
Seki S, Kohda M, Hori M (2000) Asymmetry of mouth morph of a freshwater goby, *Rhinogobius flumineus*. *Zoological Science*, **17**, 1321 1325.
Shaffer JP (1995) Multiple Hypothesis Testing. *Annual Review of Psychology*, **46**, 561 584.
Shin J, Lee C (2015) A mixed model reduces spurious genetic associations produced by population stratification in gen ome wide association studies. *Genomics*, **105**, 191 196.
Stewart TA, Albertson RC (2010) Evolution of a unique preda tory feeding apparatus: functional anatomy, development and a genetic locus for jaw laterality in Lake Tanganyika scale eating cichlids. *BMC Biology*, **8**, 11.
Sutcharit C, Asami T, Panha S (2007) Evolution of whole body enantiomorphy in the tree snail genus Amphidromus. *Journal of Evolutionary Biology*, **20**, 661 672.
Takahashi R, Moriqaki T, Hori M (2007) Foraging behaviour and functional morphology of two scale eating cichlids from Lake Tanganyika. *Journal of Fish Biology*, **70**, 1458 1469.
Takeuchi Y, Hori M (2008) Behavioural laterality in the shrimp eating cichlid fish *Neolamprologus fasciatus* in Lake Tanganyika. *Animal Behaviour*, **75**, 1359 1366.
Takeuchi Y, Hori M, Myint O, Kohda M (2010) Lateral bias of agonistic responses to mirror images and morphological asymmetry in the Siamese fighting fish (*Betta splendens*). *Behavioural Brain Research*, **208**, 106 111.
Takeuchi Y, Hori M, Oda Y (2012) Lateralized kinematics of predation behavior in a Lake Tanganyika Scale Eating Cich lid Fish. *PLoS ONE*, **7**, 10.
Takeuchi Y, Hori M, Tada S, Oda Y (2016) Acquisition of later alized predation behavior associated with development of Mouth Asymmetry in a Lake Tanganyika Scale Eating Cich lid Fish. *PLoS ONE*, **11**, e0147476.
Templeton AR (2006) *Population Genetics and Microevolutionary Theory*. John Wiley & Sons, New York.
True JR, Haag ES (2001) Developmental system drift and flexi bility in evolutionary trajectories. *Evolution & Development*, **3**, 109 119.
Turner SD (2014) qqman: an R package for visualizing GWAS results using QQ and manhattan plots. *bioRxiv*, doi:10.1101/005165.
Van Dooren TJM, van Goor HA, van Putten M (2010) Handed ness and asymmetry in scale eating cichlids: antisymmetries of different strength. *Evolution*, **64**, 2159 2165.
Wang L, Fan C, Liu Y *et al.* (2014) A Genome Scan for Quanti tative Trait Loci associated with infection resistance in Japa nese Flounder () by Bulked Segregant Analysis. *Marine Biotechnology*, **5**, 513 521.
Weir BS, Cockerham CC (1984) Estimating F statistics for the analysis of population structure. *Evolution*, **38**, 1358 1370.
Wellenreuther M, Hansonn B (2016) Detecting polygenic evolu tion: problems, pitfalls and promises. *Trends in Genetics*, **32**, 155 164.
Wigginton JE, Cutler DJ, Abecasis GR (2005) A note on exact tests of Hardy Weinberg equilibrium. *The American Journal of Human Genetics*, **76**, 887 893.
Yasugi M, Hori M (2011) Predominance of cross predation between lateral morphs in a largemouth bass and a freshwa ter goby. *Zoological Science*, **28**, 869 874.

F.R., C.F. and A.M. designed the study. Molecular analyses were performed under the supervision of P.F. F.R., P.F. and C.F. analysed the genetic data. Morphological data were collected by F.R. and analysed by C.F. F.R. and C.F. drafted the manuscript. All authors edited and agreed to the manuscript.

Data accessibility

Raw Illumina sequences and the final SNPs datasets have been archived to the NCBI's Sequence Read Archive (SRA) database with Accession no. SRA420311. The phenotypic measurements (mouth angles) and sample information (unique IDs) have been added to the DRYAD database under doi: http://dx.doi.org/10.5061/dryad.fp0b8.

Supporting information

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

Fig. S1 Enrichment bar chart. The GO terms of the sequences containing significant SNPs in the PoolSeq data set (after correcting for multiple tests with the Benjamini Hochberg (BH) procedure) are in blue. The bars corresponding to the annotation of the complete annotated gene set of *Oreochromis niloticus* (reference set) are in red.

Fig. S2 Enrichment bar chart. The GO terms of the sequences containing significant SNPs in the PoolSeq data set (after correcting for multiple tests with the Sequential Bonferroni (SB) procedure) are in blue. The bars corresponding to the annotation of the complete annotated gene set of *Oreochromis niloticus* (reference set) are in red.

Table S1 Sampling locations and sizes. The specimens from Congo were considered as one location (see main text). R right bending specimens; L left bending.

Table S2 Individual genotypes at the ddRAD locus significantly differentiated between the left and right samples after correcting for multiple tests.

Table S3 Functional annotation of the ddRAD locus with significant difference in allele frequencies between the left and right morph after correcting for multiple tests.

Table S4 AMOVA on the ddRAD data. The three random subsets and the significant SNP are reported. ** <0.05; ***<0.01.

Table S5 Functional annotation of the PoolSeq loci with significant difference in allele frequencies between the left and right morph after Benjamini Hochberg (BH) multiple tests correction.

Table S6 Functional annotation of the PoolSeq loci with significant difference in allele frequencies between the left and right morph after Sequential Bonferroni (SB) multiple tests correction.