Erschienen in: Molecular Ecology ; 26 (2017), 10. - S. 2783-2795
https://dx.doi.org/10.1111/mec.14077

quaddRAD: a new high-multiplexing and PCR duplicate removal ddRAD protocol produces novel evolutionary insights in a nonradiating cichlid lineage

PAOLO FRANCHINI DANIEL MONNÉ PARERA, ANDREAS F. KAUTT and AXEL MEYER
Zoology and Evolutionary Biology, Department of Biology, University of Konstanz, Universitätsstraße 10, 78457 Konstanz, Germany

Abstract

The identification of thousands of variants across the genomes and their accurate genotyping are crucial for estimating the genetic parameters needed to address a host of molecular ecological and evolutionary questions. With rapid advances of massively parallel high-throughput sequencing technologies, several methods have recently been developed to access genomewide data on population variation. One of the most successful and widely used techniques relies on the combination of restriction enzymes and sequencing-by-synthesis: restriction-site-associated DNA sequencing (RADSeq). We developed a new, more time- and cost-efficient double-digest RAD paired-end protocol (quaddRAD) that simplifies and speeds up the identification of PCR duplicates and permits large-scale multiplexing. Assessing its performance on a technical data set, we also applied the quaddRAD method on population samples of a Neotropical cichlid fish lineage (*Archocentrus centrarchus*) to assess its genetic structure and demographic history. While we identified allopatric interlake genetic divergence, most likely driven by drift, no signature of sympatric divergence was detected. This differs from what has been observed in the clade of Midas cichlids (*Amphilophus citrinellus* spp.), another cichlid lineage that inhabits the same lakes and shares a similar demographic history, but has evolved into small-scale adaptive radiations via sympatric speciation. We demonstrate that quaddRAD is a robust and efficient method for genotyping a massive number and widely overlapping set of loci with high accuracy. Furthermore, the results on *A. centrarchus* open new research avenues providing an ideal system to investigate genome-level mechanisms that could alter the speciation potential of different but closely related cichlid lineages.

Keywords: *Archocentrus centrarchus*, PCR duplicates, population genomics, quaddRAD, reduced representation libraries, sympatric speciation

Introduction

Only in the last decade, several methods relying on genotyping by sequencing (GBS) have revolutionized molecular ecology and evolutionary research (Andrews *et al.* 2016). These methods, combining the power of massively parallel high-throughput sequencing technologies with restriction enzyme techniques, allow the discovery and genotyping of thousands of variants across the genomes at a population level (Rowe *et al.* 2011). Despite the rapid advance of next-generation sequencing (NGS) and its associated significant reduction in cost per base, the targeting of a reduced portion of a species' genome is still a widely used strategy. Indeed, obtaining whole-genome information at a population level is still budget-prohibitive for most research groups and often unnecessary to address certain biological questions.

The opportunity to access genomewide information at a reasonable cost has made GBS techniques the method of choice to address a variety of questions in

Correspondence: Axel Meyer, Fax: +49 (0) 7531-883018;
E-mail: Axel.Meyer@uni-konstanz.de

population genetics (Hohenlohe *et al.* 2010; Kautt *et al.* 2016), phylogenetics (Jones *et al.* 2013; Eaton 2014), phylogeography (Emerson *et al.* 2010; Saenz-Agudelo *et al.* 2015) and genetic mapping (Franchini *et al.* 2014; Henning *et al.* 2014; Fruciano *et al.* 2016a) in both model and nonmodel organisms. They are particularly valuable and widely used for the latter, where screening a large number of SNPs without available genomic resources might have been cost- and time-prohibitive beforehand.

Among the GBS methods, restriction-site-associated DNA (RAD) sequencing has become the preferred method of choice in molecular ecology studies. Several RAD protocols have been published so far that mainly differ in the type and number of restriction enzymes used and in the size selection method applied (Peterson *et al.* 2012; Toonen *et al.* 2013; Schweyen *et al.* 2014; Recknagel *et al.* 2015; Tin *et al.* 2015; Hoffberg *et al.* 2016). All these variations of the original RAD protocol (Miller *et al.* 2007) have their particular advantages and limitations, as summarized by Andrews *et al.* (2016) and Puritz *et al.* (2014).

One of the most successful among the RAD approaches is the double-digest RAD (ddRAD) protocol (Peterson *et al.* 2012). Relying on two restriction enzymes and eliminating the random shearing step, ddRAD allows to recover – with appropriate size selection – a very large number of loci that are randomly distributed across the genome. Using different combinations of enzymes and varying size selections, ddRAD offers a great level of customization and makes it possible to genotype from hundreds to several thousands of orthologous loci, depending on the objectives of the case study and on the economic resources available. However, two main disadvantages can undermine the success of the ddRAD protocol. First, the implemented amplification step in the ddRAD library preparation can introduce PCR artefacts in the final sequence data set, mainly represented by PCR duplicates (Schweyen *et al.* 2014). PCR duplicates are expected to skew allele frequency estimates by increasing homozygosity, thus potentially leading to false genotype calls (Pompanon *et al.* 2005). It is therefore strongly suggested to remove PCR duplicates before genotype calling (Van der Auwera *et al.* 2013). Unfortunately, the amplification step cannot be avoided in the ddRAD method, and PCR duplicates and proper reads originating from template sequencing are usually impossible to distinguish bioinformatically – because of the nonrandom shearing step (see Peterson *et al.* 2012 for details). To overcome this limitation, the introduction of degenerate bases into the adapter sequences has been suggested as a promising approach that could enable counting the number of template molecules (Casbon *et al.* 2011; Tin *et al.* 2015).

Another disadvantage of the ddRAD protocol is the potentially variable representation of different loci in different pools/libraries that can result from uneven size selection. Machines designed for size selection of fragments (e.g. Pippin Prep, Sage Sciences), can improve this step due to their high accuracy (still 10–15% error rate) and minimize the interlibrary size variability (thus maximizing the number of orthologous loci among libraries). Yet, this source of bias can only be fully eliminated by circumventing the repeated size selection step, that is, by creating a single library pooling all individuals.

We developed a novel ddRAD paired-end protocol with the aim of (i) identifying and removing PCR duplicates by including short four-base stretches at the sequencing distal region of each Illumina adapter (P5 and P7); and (ii) increasing the sample multiplexing potential with a four barcodes strategy that incorporates two inner (one for each paired read) and two outer (one for each adapter) six-base barcodes (see Fig. 1). While removing PCR duplicates results in more accurate genotype calls, the high-multiplexing design is further beneficial in (i) reducing the cost of library preparation (many barcode combinations can be obtained using a low number of modified 'costly' oligonucleotides); (ii) minimizing the required input DNA per individual (in case many individuals are pooled); and (iii) size-selecting a single pool of hundreds of individuals in the same gel lane of either a hand-prepared electrophoresis lane or an automatic machine (thus increasing the number of overlapping loci among individuals by eliminating the random interlane size variation). Finally, our protocol is different from the original ddRAD protocol, in that restriction enzyme digestion and ligation are combined into a single reaction thereby limiting the per-sample DNA quantification steps in a streamlined protocol that reduces hands-on time compared to traditional and more recently proposed ddRAD library preparation methods. We are calling this protocol quaddRAD, where 'quadd' refers to the quadruple barcode design.

We present the method and protocol and demonstrate its utility using both a 'technical' and a 'biological' data set. For the former, 60 replicate samples were obtained from a single individual DNA extraction of a Neotropical cichlid fish for which a draft genome (Elmer *et al.* 2014) is available (*Amphilophus citrinellus*). These were then pooled to create a single quaddRAD library. Using DNA aliquots from a single sample, we were able to control for potential sources of individual variation (e.g. different level of DNA integrity; interindividual polymorphisms at restriction enzymes DNA target regions).

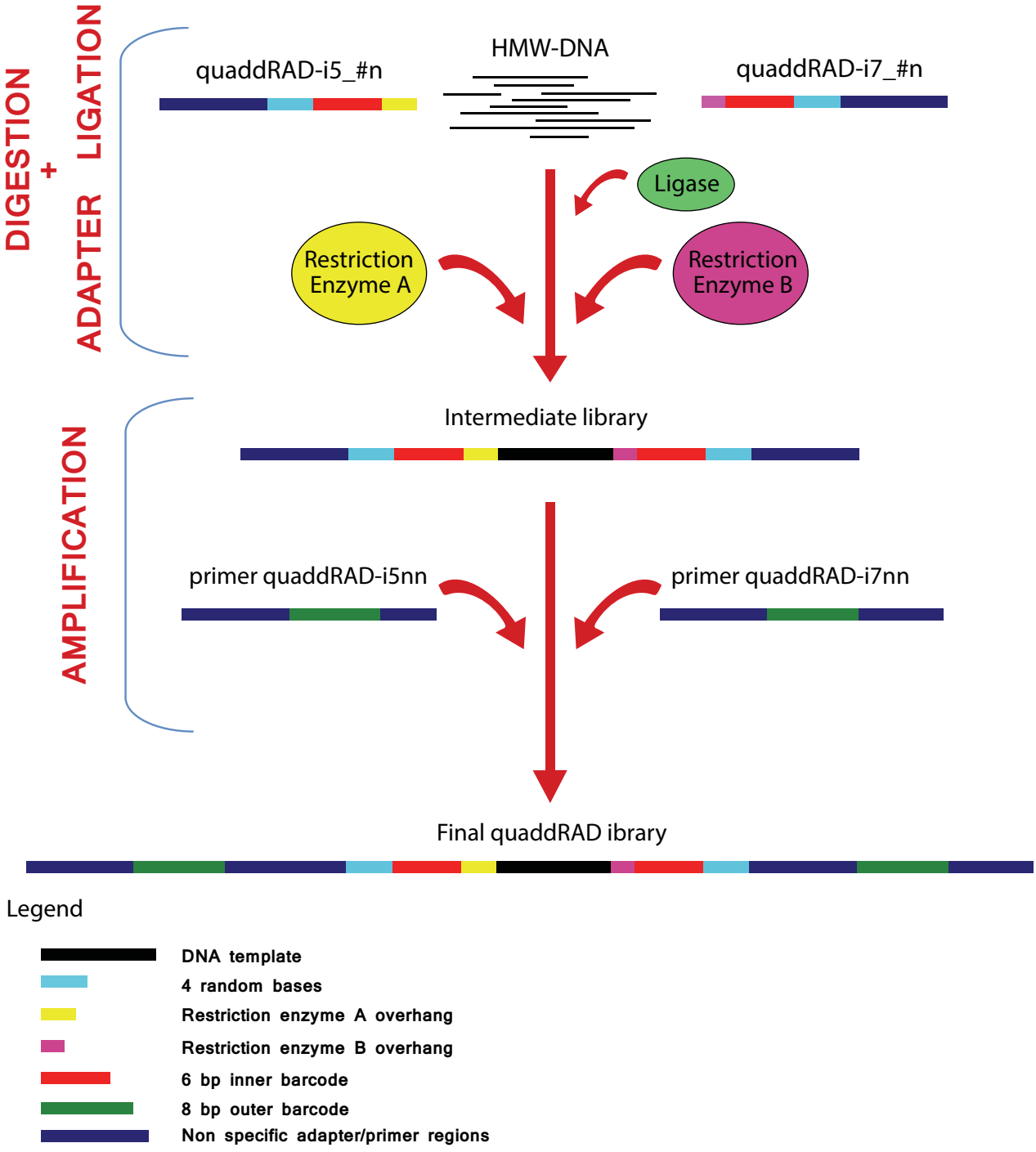

Fig. 1 Scheme of the quaddRAD workflow showing the structure of the sequencing library and the oligonucleotides that are used to construct it. The steps of the protocols are grouped when they are performed in a single reaction. [Colour figure can be viewed at wileyonlinelibrary.com]

The *A. citrinellus* species complex, fish known as Midas cichlids, mainly occur in the two large Nicaraguan lakes (Lake Nicaragua and Lake Managua), from which they have colonized a chain of crater lakes during the last few thousand years (Barluenga & Meyer 2004, 2010; Elmer *et al.* 2010). In some of these, the fish have evolved into small-scale adaptive radiations through sympatric speciation (Wilson *et al.* 2000;

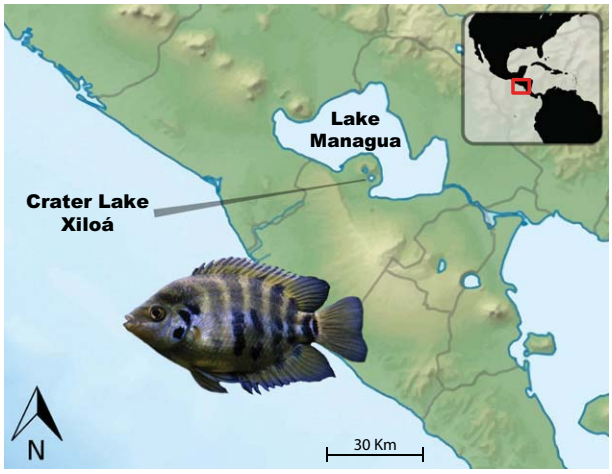

Fig. 2 Nicaraguan lake system. The *Archocentrus centrarchus* samples from this study were collected in Lake Managua and crater Lake Xiloá. A representative specimen of *A. centrarchus* is shown. [Colour figure can be viewed at wileyonlinelibrary.com]

Barluenga *et al.* 2006; Kautt *et al.* 2016). Multiple species of Midas cichlids, often showing repeated morphological differentiation, have been described within and across the lakes. Interestingly, other closely related cichlid lineages living in the same lakes have, apparently, not speciated or phenotypically diversified. Therefore in the second data set, a single quaddRAD library was constructed by pooling 67 individuals of such a monomorphic closely related Neotropical cichlid fish (*Archocentrus centrarchus*) from two freshwater basins: the large tectonic Lake Managua and the small and young crater Lake Xiloá (Fig. 2) (Fruciano *et al.* 2016b). Four endemic species of Midas cichlids have been described from Lake Xiloá and can be distinguished based on morphological (Klingenberg *et al.* 2003) and genetic clustering and these are all in turn morphologically and genetically differentiated from the Midas cichlid source population in Lake Managua (Elmer *et al.* 2010; Recknagel *et al.* 2013). We used our genomewide biological data set, first, to test whether the apparent absence of genetic clusters within the lakes for *A. centrarchus* is simply due to a lack of power, as Fruciano *et al.* (2016b) used only twelve microsatellite markers. Recently, in Lake Apoyo, another Nicaraguan crater lake hosting Midas fishes, RADSeq data comparable to those used in this study identified five genetic clusters (Kautt *et al.* 2016), while Elmer *et al.* (2014) only identified two genetic clusters using the same microsatellite set as Fruciano and colleagues.

Second, we reconstructed the demographic history of *A. centrarchus* – in a comparable manner to how it was done recently in the *A. citrinellus* species complex – to make inferences whether different demographic histories are likely to have affected the different divergence rates in the two lineages. We show that quaddRAD is an efficient, robust and cost-effective protocol for genotyping at high accuracy a massive and widely overlapping set of loci.

Materials and methods

Technical data set description

A quaddRAD library was constructed pooling 60 samples obtained from a single individual DNA extraction of a cichlid fish belonging to the Neotropical fishes of the *Amphilophus citrinellus* species complex (the same individual used for the genome assembly: (Elmer *et al.* 2014). Total DNA for this individual was extracted from fin clips using the Qiagen MagAttract extraction kit (Qiagen, Valencia, USA) and eluted in 200 μL EB buffer. Purified DNA was quantified using a Qubit 2.0 fluorimeter (Invitrogen, Carlsbad, USA) and DNA integrity was assessed by agarose gel electrophoresis. To evaluate the performance of the quaddRAD protocol we created a final pool of 60 samples in which six subpools of 10 samples each (called A, B, C, D, E and F) were prepared varying the input DNA (from 10 pg to 100 ng) and the number of PCR cycles (from 12 to 26 cycles). PCR cycles were increased from 12 to 22 for subpools A and B, respectively, while keeping the same input DNA

Table 1 Information on the six subpools of 10 samples each (technical data set) that were prepared varying the input DNA and the number of PCR cycles to evaluate the performances of the quaddRAD protocol

Input DNA (ng)	PCR cycles	Molarity factor for pooling	Illumina TruSeq outer barcodes	Total raw reads	Retained reads	Duplication rate (%)
100	12	1.00	D501-D703	30 543 232	30 420 852	0.40
100	22	1.00	D502-D704	32 605 892	32 550 094	0.17
10	15	1.00	D503-D701	41 626 610	40 264 628	3.27
1	18	1.00	D504-D702	42 352 244	32 682 470	22.83
0.1	22	0.46	D502-D703	10 230 304	7 067 388	30.92
0.01	26	0.02	D503-D702	1 146 268	785 672	31.46

(100 ng). For the other four subpools (C-F), we increased the number of PCR cycles while reducing the input DNA (Table 1). The six subpools were indexed with dual index outer barcodes (included in the Illumina adapters), while inner barcodes in both paired reads were used to index the 10 samples included in each subpool (see Table S1, Supporting information for details).

Biological data set description

A total of 67 individuals of the Neotropical cichlid species *Archocentrus centrarchus* were used in this study (Table S2, Supporting information). The fish were collected in 2012 in the great Lake Managua (23 specimens) and in the small crater Lake Xiloá (44 specimens). DNA was extracted from fin clips using the Genaxxon DNA purification kit (Genaxxon Bioscience, Ulm, Germany), and its quality and concentration were assessed using agarose gel electrophoresis and a Qubit 2.0 fluorimeter, respectively.

Adapter and primer design for quaddRAD

The protocol follows the general principles of the original ddRAD approach developed by Peterson *et al.* (2012) with modifications that allow to mark PCR duplicates, increase the multiplexing capacity and minimize hands-on time. We designed the new adapters modifying parts of the Illumina i5 and i7 adapters, here identified as quaddRAD-i5 and quaddRAD-i7, introducing in both: (i) overhangs compatible with the restriction enzymes used (3′ TGCA – PstI – in quaddRAD-i5 and 5′ GC – MspI – in quaddRAD-i7); (ii) a six-base barcode (inner barcodes); and (iii) a random four-base stretch to identify, and then bioinformatically remove, PCR duplicates. The quaddRAD-i5-top oligonucleotide was phosphorylated to block ligation at its 3′ end, while the quaddRAD-i5-bottom was 5′ phosphorylated to ligate to the DNA template (quaddRAD-i5-top: 5′-CGCTCTTC CGATCTNNNNTGCA-PHOS-3′; quaddRAD-i5-bottom: 5′-PHOS-NNNNAGATCGGAAGAGCGTCGTGTAGGG AAAGAGTGT-3′; quaddRAD-i7-top: 5′-GTGACTGGA GTTCAGACGTGTGCTCTTCCGATCTNNNN-3′; quadd RAD-i7-bottom: 5′-CGNNNNAGATCGGAAGAGCA-3′). The quaddRAD-i7 was designed as Y-shaped to ensure that only DNA fragment templates with the regular adapters on both ends will be amplified during the PCR step (Baird *et al.* 2008). All these modifications were carried out to ensure the exponential amplification of fragments with the correct adapter combination. For this study we designed three quaddRAD-i5 and four quaddRAD-i7 adapters (see Table S3, Supporting information).

After the ligation of quaddRAD-i5 and quaddRAD-i7 to our templates, the obtained fragments are enriched through a PCR step. We designed amplification primers by modifying the Illumina TruSeq i5nn and i7nn (when '*n*' indicates the barcode number) and incorporating a phosphorothioate bond at their 3′ ends to prevent unspecific/proofreading nuclease degradation. These primers (here identified as quaddRAD-i5nn: 5′-AATGA-TACGGCGACCACCGAGATCTACACACACTCTTTCC CTACACGAC*G-3′ and quaddRAD-i70n: 5′-CAAGCA-GAAGACGGCATACGAGATGTGACTGGAGTTCAGA CGTGTGC*T-3′) allow to introduce another set of two barcodes (outer barcodes) in addition to the ones introduced with the ligation of the quaddRAD-i5 and quaddRAD-i7 adapters, thus maximizing the multiplexing power by shuffling all possible combinations of the different dual index inner and outer barcodes. For this study we used 4 quaddRAD-i50n and 4 quaddRAD-i70n amplification primers (see Table S3, Supporting information).

Library preparation

A detailed version of the protocol and the samples used are provided as supplementary material (see Appendix S1, Tables S1 and S2, Supporting information). Briefly, we arranged the 127 DNA samples on two 96-well plates. Double restriction enzyme digestion (usually a rare- and a frequent-cutting enzyme is selected – here we used PstI and MspI, but any other restriction enzymes can be used depending on the target number of loci; in case different restriction enzymes are selected, the overhangs of the adapters need to be modified to match their sequences) and adapter ligation were performed in a single 40 μL reaction combining HMW genomic DNA (10 pg–1 μg), 4 μL of 10× CutSmart buffer, 0.75 μL of each restriction enzyme PstI and MspI (20 U/μL), 4 μL ATP (10 mM), 1 μL T4 DNA ligase (400 U/μL), 0.75 μL of each adapter quaddRAD-i5 and quaddRAD-i7 (10 μM) and ddH2O to reach the final volume. After incubation at 30 °C for 3 h the reaction was stopped with 10 μL EDTA (50 mM). Samples with different inner barcode combinations were then pooled. AmpureXP (Beckman Coulter, Brea, USA) beads (0.5× and 0.8×) were used to purify and size-select each pool that was eluted in 30 μL EB buffer.

In the following step, the outer barcodes were introduced by PCR combining for each pool the digested/ligated DNA (0.005–50 ng) with 20 μL 5× Phusion HF Buffer (New England Biolabs, Ipswich, USA), 2 μL dNTPs (10 mM each), 0.4 μL Phusion high-fidelity DNA polymerase (2 U/μL), 4 μL of each PCR primer quaddRAD-i5n and quaddRAD-i7n (10 μM) and ddH2O to reach 100 μL of reaction volume. PCR conditions were

as follows: 98 °C for 2 min, 10× [98 °C for 10 s, 65 °C for 30 s, 72 °C for 30 s], 72 °C for 5 min. Each amplified product, purified using AmpureXP beads (0.8×), was eluted in 20 μL EB buffer and pooled in two samples according to its belonging to the technical and the biological data set. Finally, the pooled samples were size-selected aiming for a range of 620–680 bp using a Pippin Prep electrophoresis system (Sage Science, Beverly, USA) and the final libraries were diluted to 10 nM. Paired-end Illumina next-generation sequencing (2 × 151 bp) was performed at the genomics facility of the Tufts University of Boston (TUCF Genomics) in two HiSeq 2500 lanes.

PCR duplicates removal

The raw fastq files obtained from the two Illumina lanes, demultiplexed by the sequencing provider using the outer dual index barcode information, were first processed using the *clone_filter* module, implemented in v1.35 (and subsequent versions) of the STACKS package (Catchen *et al.* 2011, 2013) to identify and remove duplicate reads. The script inspects the four bases at the beginning of each read and marks identical combinations in different paired reads. If these paired reads are also found identical for the rest of the sequence (six-base barcode and DNA template), then a single copy, stripped off the four random bases at the 5′ end of each paired reads, is retained for downstream processing. The retained sequence data set was then separated by the dual index inner barcodes (option: –inline_inline), and cleaned from erroneous and low-quality reads (options: -c –q) using the Stacks module *process_radtags*. The final filtered individuals' sequences were grouped in two data sets, one representing the technical and one the biological data set (see Tables S1 and S2, Supporting information).

Technical data set

Each one of the six libraries differing in input DNA concentration of the contained ten replicate samples was analysed separately to avoid any confounding effects. First, reads were mapped to the Midas cichlid draft reference genome (Elmer *et al.* 2014) with BWA MEM v0.7.12-r1039 (Li & Durbin 2009). Applying conservative criteria, reads with hits at several locations in the genome (identified by a –XA flag in the sam files), only local alignments (soft-clipped), or with a mapping quality of less than 25 were removed using custom bash scripts.

Loci construction and genotype calling was performed in STACKS v1.35 (Catchen *et al.* 2011, 2013) requiring a minimum of five reads to form a locus and calling genotypes at a 5% significance level using the bounded error model (upper bound of 0.05). Statistics on number of loci, utilized reads that passed all filters and were incorporated in loci, and coverage were extracted from Stacks catalogue files.

Analysis of biological data set

As no reference genome of *A. centrarchus* is available yet, the *de novo* pipeline of Stacks was applied to assemble loci with the following options: min depth of coverage to create a stack = 3, max distance allowed between stacks = 2, max distance allowed to align secondary reads = 4, max number of stacks allowed per locus = 3, deleveraging and removal algorithm = enabled.

The *rxstacks* correction module was applied in a population-specific manner (separately for individuals from Lake Managua and Lake Xiloá) to correct individual genotype calls and remove loci confounded in more than 25% of individuals or showing an excess of haplotypes. As for the technical data set, the bounded error model (upper bound of 0.05) was used and genotyping was performed at a 5% significance threshold. Afterwards, individual genotype calls with a log-likelihood of less than −100 were filtered out. Furthermore, loci showing a deviation of Hardy–Weinberg equilibrium (5% threshold) or exhibiting more than four SNPs per RAD locus were excluded. Finally, due to an excess of polymorphisms in the last three positions, loci with SNPs in these positions were excluded from subsequent analyses (see Kautt *et al.* 2016 for details).

Population clustering and demographic analyses were performed as described previously (Kautt *et al.* 2016). Briefly, using only one SNP per RAD locus and only loci present in at least ten individuals per population, population structure was investigated using ADMIXTURE v.1.23 (Alexander *et al.* 2009), principal component analyses (Patterson *et al.* 2006), and neighbour-net split graphs (Huson & Bryant 2006). The program VCFTOOLS v0.1.14 (Danecek *et al.* 2011) was used to calculate F_{ST} values at each site, while the overall genetic differentiation was calculated with ARLEQUIN v3.5.1.3 (Excoffier & Lischer 2010) and statistical significance was assessed with 10 000 permutations. Demographic inferences were based on coalescence simulations and the empirical data summarized in form of the minor (folded) site frequency spectrum (SFS) performed with FASTSIMCOAL2 v.2.5.2.3 (Excoffier *et al.* 2013). To make the estimations of genetic and demographic parameters as comparable as possible between *A. centrarchus* and *A. citrinellus* spp., we used the same version of STACKS (v1.29) used in Kautt *et al.* (2016) to avoid any bias due to different algorithms implemented in different software versions.

To account for missing data, the data for Lake Managua and Lake Xiloá were downsampled to 30 and 50 alleles (15 and 25 individuals), respectively (Gutenkunst *et al.* 2009). Five different one-population models were tested with the source population of Lake Managua and based on the results eleven different two-population models were evaluated. All models are defined and visually represented in Kautt *et al.* (2016).

Results

After sequencing we obtained a total of 502 million (M) paired sequences from two Illumina lanes with read length of 151 bp. The raw data consisted of six pools (10 individuals each) for the technical data set and six pools for the biological one (five pools with 12 individuals and 1 pool with 9), both demultiplexed according to their unique outer barcode combinations.

Technical data set

The first step of the bioinformatic workflow, intended to remove PCR duplicates, identified from a minimum of 0.17% to a maximum of 31.46% duplicated sequences in each of the six pools (called A, B, C, D, E and F), with an expected trend in which a decrease in the input DNA (and an increase in PCR cycles) resulted in a higher duplication rate (see Table 1). Out of the total 159 M, 144 M reads were retained after the PCR duplicates were removed (retained reads for each library ranged from 0.8 to 40 M; mean 23 M; SD: 1.6 M). It is important to note that the lower number of sequences obtained in pool 'E' and more drastically in pool 'F' reflect their lower concentration in the final pooling after the Pippin Prep size selection step (Table 1). The reads retained after the PCR duplicate removal step, stripped off the random four bases, were then sorted using the dual index inner barcode information and filtered by quality, generating a total of 131 M reads. The distribution of sequences for each individual sample within each of the six pools was very similar (see Tables 1 and S1, Supporting information). These sets of filtered reads, 141 bp in length after removal of the four-base random oligonucleotides and the six-base barcodes, were used for downstream analysis. Concerning the number of reads, Stacks identified a similar number of loci within each library (Table S1, Supporting information).

Biological data set: population genomics and demographic history of Archocentrus centrarchus

Across the six *A. centrarchus* pools, a duplication rate ranging from 0.59 to 1.71% was detected, allowing to retain 340 M out of the total 343 M reads (leading to a similar number of retained sequences in the six pools tagged with unique outer barcodes – from 44 to 72 M reads; mean: 56.7 M; sd: 10.1 M) (Table S4, Supporting information). After sorting the sequences by inner barcodes and trimming by quality, we obtained a total of 318 M sequences with length of 151 bp for the whole *A. centrarchus* data set of 67 individuals (from 0.9 to 8.5 M; mean: 4.6 M; sd: 1.6 M). The Stacks *de novo* pipeline detected a total of 266 995 loci in the 67 individuals (from 55 313 to 119 932 per individual; mean: 100 596; SD: 12 972) with an average coverage of 36.6× (from 10.3× to 54.8×; SD: 9.1×) (Table S2, Supporting information). After genotype calling and quality control, we identified 30 371 polymorphic sites.

To investigate whether *A. centrarchus* exhibits any so-far-undetected population structure within great Lake Managua or crater Lake Xiloá we used a set of complementary methods. These methods differ in their assumptions and using a combination of them may thus increase our sensitivity. Overall, while both lake populations are genetically clearly distinct (mean F_{ST} = 0.077, *P*-value <10e^{-5}; the distribution of the per-site F_{ST} values is shown in Fig. S1, Supporting information), none of our results suggests the presence of hidden population structure within any of the two lakes: only the first axis of the principal component analysis is significant (variance explained: PC1 = 5.52%, PC2 = 2.60%), two clusters are most supported in the Admixture analysis (i.e. had the lowest cross-validation error), and there is no visible structuring within the lakes in the individual-based split graphs (neighbour-net network) (Fig. 3). We also performed two intralacustrine (i.e. using only individuals within a lake) Admixture analyses and only one cluster was most supported in both cases (Fig. S2, Supporting information).

Among the five tested one-population demographic models, a model of population growth in Lake Managua was most supported (Table S5, Supporting information). However, except for the constant population size model, which was clearly rejected, other comparable models ('change', 'bottlegrowth') received similar support. Note that all of these models are conceptually similar and all supported a population size increase in our case. Building up on this model and in reference to models tested previously in Midas cichlids (Kautt *et al.* 2016) we tested eleven different two-population models that included both lake populations, with Lake Managua acting as source for Lake Xiloá (Table S5, Supporting information). According to the maximum-likelihood point estimates in the best model (see Table S6, Supporting information for summary and 95% confidence intervals) the population in Lake Managua started growing 3940 generations ago from ca. 32 900 individuals to a contemporary size of around 712 000

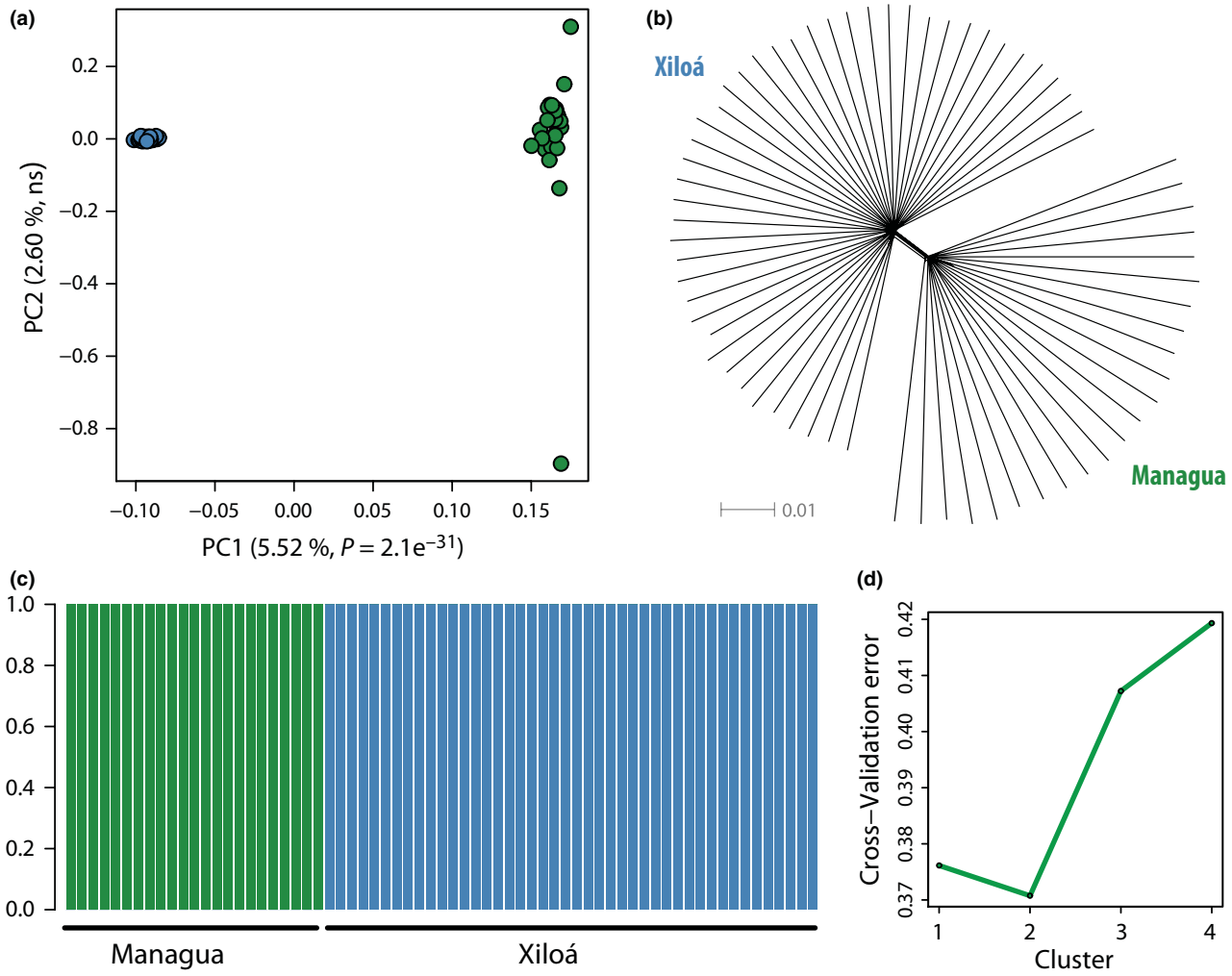

Fig. 3 Population structure analyses of the *Archocentrus centrarchus* sample: (a) principal component plot showing the distribution of the 67 individuals across the first two axes of variation; (b) individual-based neighbour-net split graph; (c) clustering analysis for individuals of the two sampled locations; and (d) associated cross-validation error graph. [Colour figure can be viewed at wileyonlinelibrary.com]

individuals. Lake Xiloá was colonized by only approximately 310 fish around 2520 generations ago and has since been growing, reaching a population size of 22 620 individuals today. In an admixture event that happened around 1300 generations ago, 56% of the gene pool in the crater lake has been replaced by the great lake. Since the colonization, there has also been continuous gene flow on the order of approximately five out of every 10 000 alleles from the great lake into the crater lake but not *vice versa* (Fig. 4).

Discussion

Here we describe a modified version of the original double-digest RAD protocol, which we term quaddRAD. We assessed its performance and demonstrate its utility in investigating the population structure and the demographic history of a Neotropical cichlid fish in the source and newly colonized crater lake population. The quaddRAD method introduces two short oligonucleotide stretches in the sequencing template that allow for the bioinformatic detection and removal of PCR duplicates, thus increasing the likelihood of correct genotype calling, especially in the case of heterozygous sites (Pompanon *et al.* 2005). Further, the four barcodes design permits a high level of multiplexing with a few tagged oligonucleotides that can be arranged in several unique combinations. The resulting pool formed by many (hundreds) individuals can be size-selected in a single electrophoretic lane, either manually or using an automatic machine, eliminating interlibrary size variability. However, it should be noted that bad quality DNA and very low concentration samples could also affect the number and distribution of sequenced loci, a

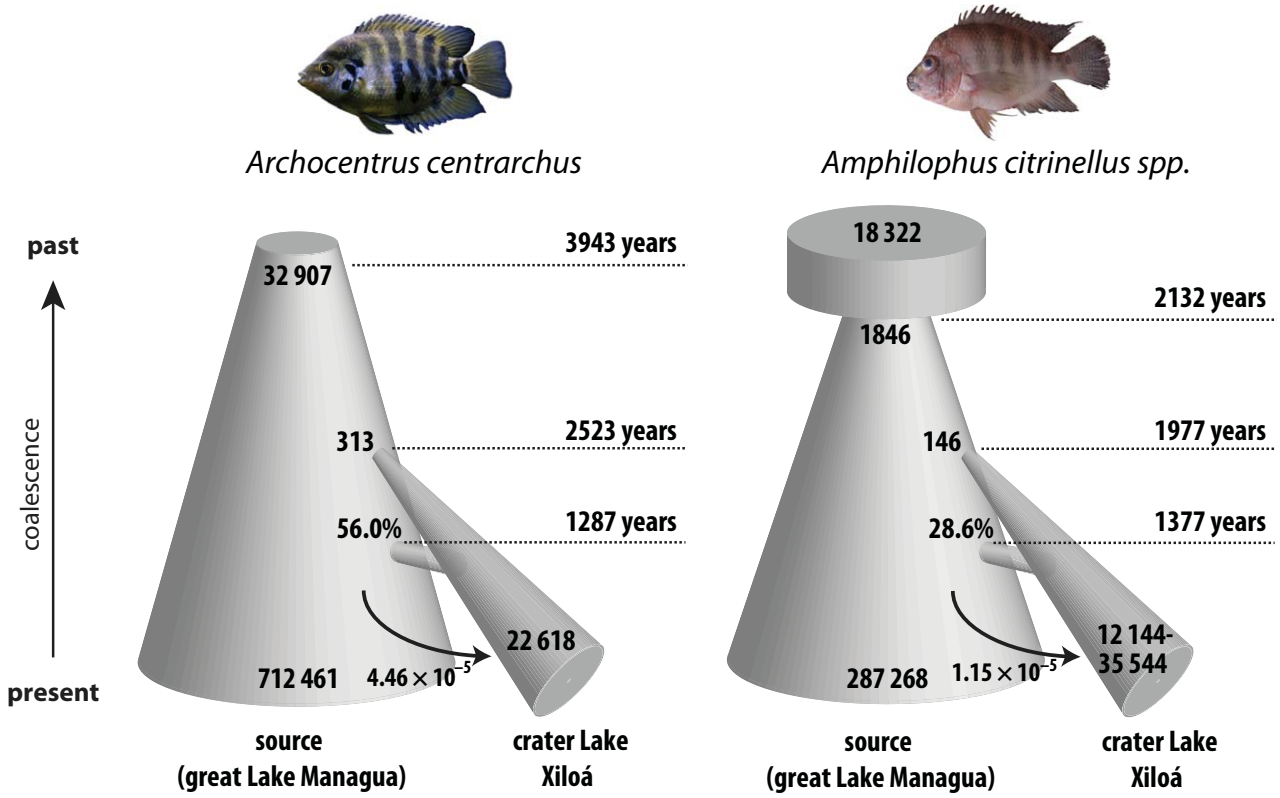

Fig. 4 Comparison of the inferred colonization history of crater Lake Xiloá by *Archocentrus centrarchus* and Midas cichlids (*Amphilophus citrinellus* species complex). Time estimates and migration rates were converted from number of generations to number of years assuming a generation time of one year for *A. centrarchus* and one and a half years for the *Amphilophus citrinellus* spp. Population sizes are given in number of individuals. Note that growth rates were modelled to be exponential and not linear as illustrated. Furthermore, models are not drawn to scale. Data for Midas cichlids are from Kautt *et al.* (2016) and the population size for Midas cichlids in Lake Xiloá gives the range of all four sympatric endemic species. [Colour figure can be viewed at wileyonlinelibrary.com]

problem that could be intensified by pooling such samples in a single reaction. These two main advantages, together with other benefits of the quaddRAD method (e.g. minimum input DNA, reduced costs of library preparation and hands-on time) make this protocol an ideal choice to be used in molecular ecology and evolutionary studies targeting large sample sizes in both model and nonmodel species.

The idea to introduce additional oligonucleotides in order to identify PCR duplicates was already suggested by Casbon *et al.* (2011) and, more recently, RAD protocols have been developed incorporating this feature (Schweyen *et al.* 2014; Tin *et al.* 2015; Hoffberg *et al.* 2016). However, our method is unique in its design as the four random bases are included in the adapters and incorporated in each paired read during the digestion/adapter ligation, the first combined step of the protocol (without including additional specific PCR cycles, thus avoiding that a proportion of the final library could miss the PCR duplication identifier). Additionally, the location of the random bases at the beginning of each read has a technical benefit. Illumina sequencing technology requires high level of base variability during the first cycles (the first five cycles/bases are the most critical) for efficient sequence cluster identification and phasing/prephasing calibration (Fadrosh *et al.* 2014). With the heterogeneous base composition of the random bases in our method, we achieve higher sequence variability during the first cycles, thus avoiding the need to spike in a certain amount of a high variable external genomic sample (e.g. PhiX). The quality and amount of the sequenced samples will then be maximized as a secondary effect of the PCR duplicate removal design.

In this study we tested the quaddRAD protocol in two cichlid fish lineages using a single combination of restriction enzymes. ddRAD protocols, which still utilize the same double-digest design of the new quaddRAD method proposed here, have been shown to be effective using a wide array of restriction enzymes in different animal taxa (e.g. Peterson *et al.* 2012; Lavretsky *et al.* 2015; Leache *et al.* 2015). Therefore, it is fair to

assume that our protocol will have the same flexibility as the ddRAD. Tests where the same individuals are genotyped using both quaddRAD and ddRAD will provide further confirmation of the proposed benefits of the new method.

Whole-genome sequencing (WGS) has recently been proposed as the ideal approach to capture the distribution of genetic variation in population samples at sufficient resolution (Seehausen *et al.* 2014; Wolf & Ellegren 2016). However, obtaining the complete genome sequence from a large number of individuals still remains prohibitively expensive and, for some specific biological questions, not even necessary. Combined with the fact that WGS is limited to species with a good quality *de novo* assembled reference genome available, this further limits its application. For these reasons, weighing advantages and limitations, techniques that reduce genome complexity will continue to be useful tools and widely used for the foreseeable future to address questions in ecological and evolutionary research, especially in nonmodel species.

The Midas/non-Midas mystery

The small adaptive radiations of the *Amphilophus citrinellus* species flock from the Nicaraguan lakes have diversified in several phenotypic axes such as body shape, coloration and feeding apparatus (Barluenga *et al.* 2006; Elmer *et al.* 2014; Franchini *et al.* 2016; Fruciano *et al.* 2016a). By contrast, no clear phenotypic diversification has been observed in any other cichlid lineage that is inhabiting the same lake system. But we note that the Midas system has been the focus of a huge amount of research in comparison to the other non-Midas lineages.

Recently, Fruciano *et al.* (2016b) did not find any evidence of morphological (both within and between lakes) and genetic (within lakes) divergence in the cichlid fish *Archocentrus centrarchus*, and showed no significant difference in timing of colonization of crater Lake Xiloá from Lake Managua between the *A. citrinellus* species flock and *A. centrarchus*. Fruciano *et al.* used twelve microsatellite loci to assess population structure and the mitochondrial control region for the demographic analyses. Here we used quaddRAD to improve the accuracy of the molecular surveys with a genome-wide approach using thousands of markers. We intended to make this data set comparable with that of our earlier study (Kautt *et al.* 2016) that was based on an earlier ddRAD protocol. Only low, but strongly significant, genetic differentiation between lakes was found (due to the contribution of the majority of loci), with all individuals being assigned to their lake of origin without any signature of recent admixture. In contrast, using microsatellite markers, Fruciano *et al.* (2016b) found a pattern that could have been interpreted as admixture in a few individuals – this interpretation was possibly due to a lack of power/resolution. Consistent with the results from Fruciano *et al.* (2016b), even with our high-resolution genomewide data set we did not detect any genetic clustering within Lake Xiloá, suggesting that there is no sympatric differentiation in *A. centrarchus*. Conversely, Kautt *et al.* (2016) detected distinguishable genetic clusters within crater Lake Xiloá in the *A. citrinellus* species flock (well matching the four described endemic species) and suggested that these species have evolved through sympatric speciation after a second wave of colonization from Lake Managua, that possibly boosted the speciation process.

The demographic history that we inferred for *A. centrarchus* mirrored the one of the *A. citrinellus* spp., except for the fact that we did not find evidence for a bottleneck in the great lake population. Indeed, the most supported model was almost identical to the best model in Midas cichlids (Fig. 4). The lack of a signal for a bottleneck in the source population of Managua could be due to a lack of power as our site frequency spectrum (SFS) for *A. centrarchus* was smaller than the one found for the *A. citrinellus* spp., with 30 vs. 50 alleles, respectively (Robinson *et al.* 2014). The maximum-likelihood parameter estimates of the population sizes, gene flow, and the admixture event are overall similar between the *A. citrinellus* species complex and *A. centrarchus* and were, except for the ancestral size of Lake Managua, reciprocally contained within the 95% confidence intervals (Table S6, Supporting information).

Time points between the models are more difficult to compare, as they are in units of generations and there is some uncertainty in the generation times of the two lineages. Kautt *et al.* (2016) assumed a generation time of one to two years in the wild for *A. citrinellus*. Given that *A. centrarchus* are somewhat smaller than *A. citrinellus* and reproduce earlier in the laboratory (personal observation) it is probably fair to assume that they have a shorter generation time. If we assume the generation time of *A. centrarchus* to be around two-thirds of the one of *A. citrinellus* (i.e. one year for *A. centrarchus* and one and a half years for *A. citrinellus*) the estimated time points of colonization of Lake Xiloá translate – interestingly – to a very similar time for both lineages. While absolute estimates obtained from two different data sets are potentially prone to differences for technical reasons (see Kautt *et al.* 2016) for discussion of factors that could influence estimates) we suggest that the demographic histories between the two lineages are very similar. Under this scenario, both *A. citrinellus* spp. and *A. centrarchus* colonized a new environment around 2000–2500 years ago that offered them the same ecological opportunities, but only the

former one diverged from the source population and radiated *in situ*. Why *A. centrarchus* did not speciate while the *A. citrinellus* species complex living in the same lakes did remains an open question and at this stage we can only speculate on the reasons. Disruptive selection, for example, could have been stronger in the *A. citrinellus* spp., thereby facilitating a faster divergence of adaptive traits accompanied by the occupation of the available ecological niches. Another possible explanation is that, even if disruptive selection promoted a first (weak) phenotypic divergence in the two lineages, *A. centrarchus* could not evolve strong assortative mating (that has been recently documented in Midas: Elmer *et al.* 2009; Machado-Schiaffino *et al.* 2017). Traits with potential adaptive values in the two lineages could have different underlying genetic bases that could alternatively hinder or promote their divergence even under similar selective regimes (Fruciano *et al.* 2016a). Therefore, a more thorough investigation of the genomic differences between the two lineages would be one obvious next step to shed new insight on their apparent different potential or propensity for speciation. In this regard, it has recently been shown how structural variation can have a large impact on genome evolution and, therefore, we aim to consider this genomic aspect in future studies on the genetics of adaptation and speciation (Chain & Feulner 2014; Feder *et al.* 2014). To determine the presence and extent of structural variations between the two focal lineages, we aim for sequencing of the *A. centrarchus* genome that, together with individual whole-genome resequencing of population samples of both *A. centrarchus* and the *A. citrinellus* species complex from different lakes, will represent valuable resources to further foster high-resolution comparative genomic studies.

Acknowledgements

We thank C. Fruciano and G. Machado-Schiaffino for their help in fish sampling, B. Rueter for laboratory assistance and E. Malecore for graphical support. In particular, we thank Julian Catchen for earlier discussions of our work and for developing the PCR duplicate detection script that is available from version 1.35 of the STACKS package. This work was supported by the ERC Advanced grant GenAdap 293700 by the European Research Council to AM. PF was supported by a Deutsche Forschungsgemeinschaft Research Grant (FR 3399/1-1). The University of Konstanz and the GeCKo (University of Konstanz Genomic Center) are thanked for their support.

References

Alexander DH, Novembre J, Lange K (2009) Fast model-based estimation of ancestry in unrelated individuals. *Genome Research*, **19**, 1655–1664.

Andrews KR, Good JM, Miller MR, Luikart G, Hohenlohe PA (2016) Harnessing the power of RADseq for ecological and evolutionary genomics. *Nature Reviews Genetics*, **17**, 81–92.

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

Barluenga M, Meyer A (2004) The Midas cichlid species complex: incipient sympatric speciation in Nicaraguan cichlid fishes? *Molecular Ecology*, **13**, 2061–2076.

Barluenga M, Meyer A (2010) Phylogeography, colonization and population history of the Midas cichlid species complex (*Amphilophus* spp.) in the Nicaraguan crater lakes. *BMC Evolutionary Biology*, **10**, 326.

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

Casbon JA, Osborne RJ, Brenner S, Lichtenstein CP (2011) A method for counting PCR template molecules with application to next-generation sequencing. *Nucleic Acids Research*, **39**, e81.

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.

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

Chain FJ, Feulner PG (2014) Ecological and evolutionary implications of genomic structural variations. *Frontiers in Genetics*, **5**, 326.

Danecek P, Auton A, Abecasis G *et al.* (2011) The variant call format and VCFtools. *Bioinformatics*, **27**, 2156–2158.

Eaton DAR (2014) PyRAD: assembly of de novo RADseq loci for phylogenetic analyses. *Bioinformatics*, **30**, 1844–1849.

Elmer KR, Lehtonen TK, Meyer A (2009) Color assortative mating contributes to sympatric divergence of neotropical cichlid fish. *Evolution*, **63**, 2750–2757.

Elmer KR, Kusche H, Lehtonen TK, Meyer A (2010) Local variation and parallel evolution: morphological and genetic diversity across a species complex of neotropical crater lake cichlid fishes. *Philosophical Transactions of the Royal Society B-Biological Sciences*, **365**, 1763–1782.

Elmer KR, Fan S, Kusche H *et al.* (2014) Parallel evolution of Nicaraguan crater lake cichlid fishes via non-parallel routes. *Nature Communications*, **5**, 5168.

Emerson KJ, Merz CR, Catchen JM *et al.* (2010) Resolving postglacial phylogeography using high-throughput sequencing. *Proceedings of the National Academy of Sciences of the United States of America*, **107**, 16196–16200.

Excoffier L, Lischer HEL (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, Dupanloup I, Huerta-Sanchez E, Sousa VC, Foll M (2013) Robust demographic inference from genomic and SNP data. *PLoS Genetics*, **9**, 10.

Fadrosh DW, Ma B, Gajer P *et al.* (2014) An improved dual-indexing approach for multiplexed 16S rRNA gene sequencing on the Illumina MiSeq platform. *Microbiome*, **2**, 6.

Feder JL, Nosil P, Flaxman SM (2014) Assessing when chromosomal rearrangements affect the dynamics of speciation: implications from computer simulations. *Frontiers in Genetics*, **5**, 295.

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.

Franchini P, Xiong P, Fruciano C, Meyer A (2016) The role of microRNAs in the repeated parallel diversification of lineages of Midas cichlid fish from Nicaragua. *Genome Biology and Evolution*, **8**, 1543–1555.

Fruciano C, Franchini P, Kovacova V *et al.* (2016a) Genetic linkage of distinct adaptive traits in sympatrically speciating Crater Lake cichlid fish. *Nature Communications*, **7**, 12736.

Fruciano C, Franchini P, Raffini F, Fan S, Meyer A (2016b) Are sympatrically speciating Midas cichlid fish special? Patterns of morphological and genetic variation in the closely related species *Archocentrus centrarchus*. *Ecology and Evolution*, **6**, 4102–4114.

Gutenkunst RN, Hernandez RD, Williamson SH, Bustamante CD (2009) Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS Genetics*, **5**, 10.

Henning F, Lee HJ, Franchini P, Meyer A (2014) Genetic mapping of horizontal stripes in Lake Victoria cichlid fishes: benefits and pitfalls of using RAD markers for dense linkage mapping. *Molecular Ecology*, **23**, 5224–5240.

Hoffberg SL, Kieran TJ, Catchen JM *et al.* (2016) RADcap: sequence capture of dual-digest RADseq libraries with identifiable duplicates and reduced missing data. *Molecular Ecology Resources*, **16**, 1264–1278.

Hohenlohe PA, Bassham S, Etter PD, *et al.* (2010) Population genomics of parallel adaptation in threespine stickleback using sequenced RAD tags. *Plos Genetics*, **6**, e1000862.

Huson DH, Bryant D (2006) Application of phylogenetic networks in evolutionary studies. *Molecular Biology and Evolution*, **23**, 254–267.

Jones JC, Fan SH, Franchini P, Schartl M, Meyer A (2013) The evolutionary history of Xiphophorus fish and their sexually selected sword: a genome-wide approach using restriction site-associated DNA sequencing. *Molecular Ecology*, **22**, 2986–3001.

Kautt AF, Machado-Schiaffino G, Meyer A (2016) Multispecies outcomes of sympatric speciation after admixture with the source population in two radiations of Nicaraguan Crater Lake cichlids. *Plos Genetics*, **12**, e1006157.

Klingenberg CP, Barluenga M, Meyer A (2003) Body shape variation in cichlid fishes of the *Amphilophus citrinellus* species complex. *Biological Journal of the Linnean Society*, **80**, 397–408.

Lavretsky P, Dacosta JM, Hernandez-Banos BE *et al.* (2015) Speciation genomics and a role for the Z chromosome in the early stages of divergence between Mexican ducks and mallards. *Molecular Ecology*, **24**, 5364–5378.

Leache AD, Chavez AS, Jones LN *et al.* (2015) Phylogenomics of phrynosomatid lizards: conflicting signals from sequence capture versus restriction site associated DNA sequencing. *Genome Biology and Evolution*, **7**, 706–719.

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

Machado-Schiaffino G, Kautt AF, Torres-Dowdall J *et al.* (2017) Incipient speciation driven by hypertrophied lips in Midas cichlids fish? *Molecular Ecology*, **26**, 2348–2362.

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.

Patterson N, Price AL, Reich D (2006) Population structure and eigenanalysis. *PLoS Genetics*, **2**, 2074–2093.

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.

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

Puritz JB, Matz MV, Toonen RJ *et al.* (2014) Demystifying the RAD fad. *Molecular Ecology*, **23**, 5937–5942.

Recknagel H, Kusche H, Elmer KR, Meyer A (2013) Two new endemic species in the Midas cichlid species complex from Nicaraguan crater lakes: *Amphilophus tolteca* and *Amphilophus viridis* (Perciformes, Cichlidae). *Aqua*, **19**, 207–224.

Recknagel H, Jacobs A, Herzyk P, Elmer KR (2015) Double-digest RAD sequencing using Ion Proton semiconductor platform (ddRADseq-ion) with nonmodel organisms. *Molecular Ecology Resources*, **15**, 1316–1329.

Robinson JD, Coffman AJ, Hickerson MJ, Gutenkunst RN (2014) Sampling strategies for frequency spectrum-based population genomic inference. *Bmc Evolutionary Biology*, **14**, 254.

Rowe HC, Renaut S, Guggisberg A (2011) RAD in the realm of next-generation sequencing technologies. *Molecular Ecology*, **20**, 3499–3502.

Saenz-Agudelo P, Dibattista JD, Piatek MJ *et al.* (2015) Seascape genetics along environmental gradients in the Arabian Peninsula: insights from ddRAD sequencing of anemonefishes. *Molecular Ecology*, **24**, 6241–6255.

Schweyen H, Rozenberg A, Leese F (2014) Detection and removal of PCR duplicates in population Genomic ddRAD studies by addition of a degenerate base region (DBR) in sequencing adapters. *Biological Bulletin*, **227**, 146–160.

Seehausen O, Butlin RK, Keller I *et al.* (2014) Genomics and the origin of species. *Nature Reviews Genetics*, **15**, 176–192.

Tin MMY, Rheindt FE, Cros E, Mikheyev AS (2015) Degenerate adaptor sequences for detecting PCR duplicates in reduced representation sequencing data improve genotype calling accuracy. *Molecular Ecology Resources*, **15**, 329–336.

Toonen RJ, Puritz JB, Forsman ZH *et al.* (2013) ezRAD: a simplified method for genomic genotyping in non-model organisms. *Peerj*, **1**, e203.

Van der Auwera GA, Carneiro MO, Hartl C *et al.* (2013) From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Current Protocols in Bioinformatics*, **43**, 11 10 11-33.

Wilson AB, Noack-Kunnmann K, Meyer A (2000) Incipient speciation in sympatric Nicaraguan crater lake cichlid fishes: sexual selection versus ecological diversification. *Proceedings of the Royal Society B-Biological Sciences*, **267**, 2133–2141.

Wolf JBW, Ellegren H (2016) Making sense of genomic islands of differentiation in light of speciation. *Nature Reviews Genetics*, advance online publication.

P.F. conducted molecular data analyses and wrote the manuscript. D.M.P. carried out the laboratory work. P.F. and D.M.P. conceived the study. A.F.K. conducted molecular data analyses and wrote the manuscript. A.M. wrote the manuscript.

Fig. S1 Distribution of F_{ST} values at all 30 371 polymorphic sites.

Fig. S2 Cross-validation error scores for the two intralacustrine *Admixture* analyses.

Table S1 For each individual sample, barcode combination used, total number of raw and filtered reads, number of RAD loci identified using the *ref_map.pl* Stacks pipeline and their mean depth of coverage are shown (see Materials and Methods for details on the parameters used in Stacks).

Table S2 For each *Archocentrus centrarchus* individual, lake of origin, barcode combination used, total number of raw and filtered reads, number of RAD loci identified using the *ref_map.pl* Stacks pipeline and their mean depth of coverage are shown (see Materials and Methods for details on the parameters used in Stacks).

Table S3 Oligonucleotide sequences of the adapters and primers that we used in this study.

Table S4 Information on the 67 *Archocentrus centrarchus* individuals (biological data set), where the pooling design and the statistics about number of reads and RAD loci are shown.

Table S5 Support for tested one- and two-population demographic models.

Table S6 Summary of parameter estimates in most supported demographic model.

Appendix S1 Complete laboratory protocol for the execution of the quaddRAD method.