MOLECULAR ECOLOGY

Molecular Ecology (2013) **22**, 2986–3001 doi: 10.1111/mec.12269

The evolutionary history of *Xiphophorus* fish and their sexually selected sword: a genome-wide approach using restriction site-associated DNA sequencing

JULIA C. JONES,*† SHAOHUA FAN,* PAOLO FRANCHINI,* MANFRED SCHARTL‡ and AXEL MEYER*

*Lehrstuhl für Zoologie und Evolutionsbiologie, Department of Biology, University of Konstanz, Universitätsstrasße 10, 78457 Konstanz, Germany, †Zukunftskolleg, University of Konstanz, Konstanz, Germany, ‡Physiological Chemistry, Biozentrum, University of Würzburg, Am Hubland, 97074 Würzburg, Germany

Abstract

Next-generation sequencing (NGS) techniques are now key tools in the detection of population genomic and gene expression differences in a large array of organisms. However, so far few studies have utilized such data for phylogenetic estimations. Here, we use NGS data obtained from genome-wide restriction site-associated DNA (RAD) (~66000 SNPs) to estimate the phylogenetic relationships among all 26 species of swordtail and platyfish (genus *Xiphophorus*) from Central America. Past studies, both sequence and morphology-based, have differed in their inferences of the evolutionary relationships within this genus, particularly at the species-level and among monophyletic groupings. We show that using a large number of markers throughout the genome, we are able to infer the phylogenetic relationships with unparalleled resolution for this genus. The relationships among all three major clades and species within each of them are highly resolved and consistent under maximum likelihood, Bayesian inference and maximum parsimony. However, we also highlight the current cautions with this data type and analyses. This genus exhibits a particularly interesting evolutionary history where at least two species may have arisen through hybridization events. Here, we are able to infer the paternal lineages of these putative hybrid species. Using the RAD-marker-based tree we reconstruct the evolutionary history of the sexually selected sword trait and show that it may have been present in the common ancestor of the genus. Together our results highlight the outstanding capacity that RAD sequencing data has for resolving previously problematic phylogenetic relationships, particularly among relatively closely related species.

Keywords: hybridization, phylogenetics, pre-existing bias hypothesis, RAD sequencing, sexual selection *Xiphophorus*

Received 30 June 2012; revision received 10 January 2013; accepted 15 January 2013

Introduction

High-throughput next-generation sequencing (NGS) technologies are now key tools in the rapid detection of genetic variation among both model and nonmodel organisms. These methods have transformed the questions that can be addressed and the taxa that can be studied using genome-wide approaches. To date, however, few studies have applied this technology to phylogenetics (but see Emerson *et al.* 2010; Dasmahapatra *et al.* 2012; Rubin *et al.* 2012; Wagner *et al.* 2012). As the costs are reduced further and the use of genome-wide markers and associated analyses becomes more and more efficient, there will be greater possibilities for investigating a wide range of taxa.

Specifically, Baird *et al.* (2008) developed a new method for sampling genome-wide SNP variation in

Correspondence: Axel Meyer, Fax: +49 7531 883018;
E-mail: axel.meyer@uni-konstanz.de

restriction site-associated DNA (termed RAD) using short-read NGS technology (e.g. Illumina and SOLiD NGS platforms). This method targets a reduced representation of the genome—orthologous regions flanking restriction enzyme cutting sites throughout the genome. To date, RAD sequencing has enabled the production of genome-wide SNP data for the use in population genomic studies of threespine sticklebacks (Hohenlohe *et al.* 2010, 2012) and hybridization between rainbow and westslope cutthroat trout (Hohenlohe *et al.* 2011). In addition, Emerson *et al.* (2010) used RAD sequencing to identify the previously difficult to resolve phylogeographic relationships among populations of the pitcher plant mosquito (*Wyeomyia smithii*). Dasmahapatra *et al.* (2012) used RAD sequencing to reconstruct a phylogenetic tree of species in the *melpomene*–silvaniform clade of *Heliconius* butterflies. Rubin *et al.* (2012) focussed on assessing whether RAD sequences (simulated) can be used across a broad range of species (i.e. yeast, *Drosophila* and mammals) without a reference genome and understanding which parameters yield the most accurate and well-supported trees. This method has also been applied very recently to resolving the evolutionary relationships in the Lake Victoria cichlid adaptive radiation (Wagner *et al.* 2012). Here, we use RAD sequencing to investigate the phylogenomic relationships among all species of swordtail and platyfish (genus *Xiphophorus*) from Mexico, South America, and test the capacity of this type of data for providing high-resolution estimates of species relationships. We also outline some of the current potential limits and biases of applying phylogenetic analyses to RAD data. Challenges associated with genome-wide scale phylogenetic analyses are starting to be uncovered (e.g. Kumar *et al.* 2012; Simmons 2012a,b); however, a more extensive use of these methods will provide better insight into the broader patterns of phylogenetic inference based on such large data sets.

The freshwater fish genus *Xiphophorus* (swordtails and platyfish) belongs to the Family Poeciliidae and includes 26 species of small fish from Central America (Kallman & Kazianis 2006). This group of fish has been widely used as a model for a range of evolutionary and ecological questions including mating preferences and asymmetries, fitness differences and conservation genetics and the evolution of unisexual populations and provides remarkable opportunities for genomic studies of behavioural and ecological radiations and speciation (e.g. Vrijenhoek *et al.* 1985; Ryan & Wagner 1987; Quattro & Vrijenhoek 1989; Basolo 1990b, 1998; Kirkpatrick & Ryan 1991; Quattro *et al.* 1996; Stöck *et al.* 2010; Willing *et al.* 2010; McCoy *et al.* 2011; Shen *et al.* 2012). *Xiphophorus* fish in particular are also models in cancer (melanoma) research, including the relationship between oncogenes and speciation (e.g. see Schartl 2008). In addition, these fish have been the focus of much research on the molecular mechanisms driving the evolution of sex determination (e.g. Volff & Schartl 2001; Schartl 2004), and interestingly, gene copy number has been found to be associated with reproductive strategy, size and puberty (Lampert *et al.* 2010). *Xiphophorus* has perhaps attracted the most research attention for their sexually selected male trait—the elongated ventral caudal fin or sword (e.g. Meyer *et al.* 1994, 2006; Basolo 1995a,b; Rosenthal & Evans 1998; Rosenthal *et al.* 2002). Swordtail males exhibit a sword trait that is preferred by females, and even female platyfish whose males do not have a sword prefer conspecific males with artificial swords over swordless ones (Basolo 1990a). *Xiphophorus* have been suggested to be one of few examples of the pre-existing bias hypothesis, where evolutionary older female mating preferences for sworded males may have driven the evolution of the more recently evolved sword trait (Basolo 1995a,b; Meyer 1997).

Estimating the phylogenetic relationships among this group of fish has been much addressed but has provided conflicting results in the literature with different data sets illustrating different evolutionary scenarios and monophyletic groupings (Rosen 1960, 1979; Rosen & Kallman 1969; Rauchenberger *et al.* 1990; Basolo 1991; Meyer *et al.* 1994, 2006). Until the mid-1990s, phylogenetic relationships among *Xiphophorus* fish were estimated using morphological traits such as the sword, pigmentation and the elaborate fin-structure of the gonopodium, which is the intromittent organ that develops from the anal fin of males in these live bearing fishes (Rosen 1960, 1979; Rosen & Kallman 1969; Rauchenberger *et al.* 1990; Basolo 1991). Since then, molecular markers, both mtDNA and nuclear, have been used to estimate the relationships among these fish (e.g. Meyer *et al.* 1994, 2006; Kang *et al.* 2013). Most phylogenies support four clades within this genus—southern swordtails (not always resolved as monophyletic), northern swordtails, southern platyfish and northern platyfish. However, more derived nodes within these clades remain not well supported.

Of particular interest are the putative signatures of hybridization in this group, revealed through incongruence between mitochondrial and nuclear genetic markers (Meyer *et al.* 1994, 2006; Kang *et al.* 2013). However, we note that such signatures may arise for a number of different reasons, including differences in the evolutionary processes of the nuclear vs. mitochondrial genome (Ballard & Whitlock 2004). At least two species of *Xiphophorus* show discordance in their phylogenetic placement and have been suggested to have arisen through hybridization events (Meyer *et al.* 1994, 2006; Jones *et al.* 2012; Kang *et al.* 2013).

Here, we use SNP data from thousands of RAD loci to provide high-resolution estimates of the evolutionary relationships among all 26 *Xiphophorus* fish species. We also estimate the divergence times of the broader groupings of this genus and the origin of the putative hybrid species. Additionally, we address the question of which regions of the genome are shared between fish likely to have been affected by introgression. Together, our results show that RAD sequencing can provide exceptional insights into the evolutionary relationships among an entire genus.

Methods

Samples

We sequenced an average of five individuals of each of the 26 described species of *Xiphophorus*, plus three outgroups (*Priapella intermedia*, *Gambusia holbrooki*, *Heterandria formosa*) (sample numbers per species ranged from 2 to 7), making a total of 143 individuals (Table 1). All samples, apart from *X. malinche* that were wild-caught, were from laboratory-housed stocks. All stocks are derived from wild-caught fish (~5–20 founders) and have been maintained under closed colony breeding in large population tanks for a range of 5–50 generations. For each species stock, at least two independent aquaria are kept, and random mixing of individuals is performed every 2–3 generations.

Molecular methods

DNA was extracted from body tissue using a DNeasy Blood and Tissue kit (Qiagen, Valencia, CA, USA) with RNase A (100 mg/mL). DNA was quantified using a TBS-380 Mini-Fluorometer (Turner Biosystems, Sunnyvale, CA, USA), and the quality was assessed by visualization on an agarose gel.

Restriction site-associated DNA libraries were prepared following the study by Peterson *et al.* (2012) with modifications as in the study by Recknagel *et al.* (2013) and described here. Briefly, 500 ng of DNA from each individual was digested using two enzymes, one rare cutting enzyme (*Pst*I-HF) and one frequent cutting enzyme (*Msp*I), for 3 h at 37 °C. The fragmented DNA was purified using the Qiagen MinElute PCR Purifica-

Table 1 Specimens by origin and species. Individuals sampled here are from laboratory strains bred from wild-caught individuals from the localities specified

Species	Origin	GPS location	*N* per species
Xiphophorus andersi	Río Atoyac		5
Xiphophorus alvarezi	Rio Dolores		5
Xiphophorus birchmanni	Río Axtlapexco	N 21°02′13,5″ W 98°22′22,5″	5
Xiphophorus clemenciae	Puente Chino Luiz		5
Xiphophorus continens	Ojo Frío	N 22°11.432′ W 99°19.326′	5
Xiphophorus cortezi	Río Axtla		5
Xiphophorus couchianus	Apodaca		5
Xiphophorus evelynae	Tecolutla		5
Xiphophorus gordoni	Cuatro Cienagas		5
Xiphophorus hellerii	Río Lancetilla		5
Xiphophorus kallmani	Laguna Catemaco	N 18°22.256′ W 95°00.057′	5
Xiphophorus malinche	Arroyo Xontla, near Chicayotla	N 20°55′ 26″ W 98°34′ 35″	5
Xiphophorus maculatus	Río Grijalva		5
Xiphophorus mayae	Río Dulce		5
Xiphophorus meyeri	Melchor Musquiz		5
Xiphophorus milleri	Laguna Catemaco		5
Xiphophorus mixei	Rio del Sol		2
Xiphophorus montezumae	Cascadas de Tamasopo	N 21°56.411′ W 99°23.703′	5
Xiphophorus monticolus	El Tejon		7
Xiphophorus multilineatus	Rio Coy	N 21°45.096′ W 98°57.445′	5
Xiphophorus nezahualcoyotl	Río El Salto		5
Xiphophorus nigrensis	Nacimiento de Choy	N 21°59.264′ W 98°53.106′	5
Xiphophorus pygmaeus	Río Axtla		5
Xiphophorus signum	Río Chajmaic		4
Xiphophorus variatus	Cuidad Mante	N 22°48′43.8″ W99°00′45.2″	5
Xiphophorus xiphidium	Río Purificación	N 24°02.848′ W 98°22.264′	5
Gambusia holbrooki	Everglades, Florida		5
Heterandria formosa	Everglades, Florida		5
Priapella intermedia	Río La Lana	N 19°02.536′ W 96°10.471	5

tion Kit protocol, eluted in 10 μL EB (×2) and again quantified. Next, unique P1 adapters (i.e. see https://www.wiki.ed.ac.uk/display/RADSequencing/Home) that bind to *Pst*I-HF created restriction sites were ligated to DNA fragments from each individual, and P2 adapters were ligated to sites created by *Msp*I (Fwd: CGAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGATCAGAACAA, Rev: CAAGCAGAAGACGGCATACGAGATCGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATCT). Each reaction consisted of a combination of ~400 ng of DNA, 1 μL of P1 adapter (10 μM), 1 μL of P2 adapter (10 μM), 1 μL T4 ligase (1,000 U/μL), 4 μL of 10× T4 ligation buffer and ddH_2O to a total volume of 40 μL. The ligation step was completed on a PCR machine using the following conditions: 25 °C for 30 min, 65 °C for 10 min and then the temperature was decreased to 20 °C at 1.3 °C per minute. After ligation, samples were again purified using the Qiagen MinElute PCR Purification Kit, then multiplexed and manually size-selected using gel electrophoresis (300–400 bp). Appropriately sized bands were cut from the gel with sterilized razor blades and cleaned using the Qiagen MinElute Gel Purification kit, and eluted in 10 μL EB (×2).

The DNA concentration of each multiplexed sample was measured and then amplified in eight single PCRs per library. Each PCR contained 10–20 ng of library DNA template, 4 μL dNTPs (100 mM), 4.0 μL 5× Phusion HF buffer (NEB), 0.2 μL Phusion *Taq* polymerase (NEB) and 1.0 μL of each RAD primer (10 μM), made up to 20 μL with ddH_2O. The PCR conditions for library amplification were 98 °C (30 s), [98 °C (10 s), 65 °C (30 s), 72 °C (30 s)] × 10, 72 °C (300 s). PCR products for each library were combined and cleaned using gel electrophoresis. Gel bands (300–400 bp) were cut, again cleaned using the Qiagen MinElute Gel Extraction Kit and eluted in 10 μL buffer EB. Two of the three libraries were sequenced in an Illumina Genome Analyzer IIx, two lanes per library, single-end sequencing, yielding a maximum read length of 150 bp. The third library was sequenced in an Illumina HiSeq 2000, one lane, single-end sequencing, yielding maximum read lengths of 100 bp (see Table S1, Supporting information, for how samples were multiplexed).

Raw read quality filtering and processing

Raw sequence reads were processed using the Stacks pipeline (v0.997, process_radtags) (Catchen *et al.* 2011). The parameters (–t 100 -r -q -c -E –b barcode) were specified in order to truncate all the raw reads to 100 bp in length, rescue the ambiguous sites in the barcode sites (because the barcodes used in this study differ by at least two base pairs between samples, reads with ambiguous sites in the barcode region could be first corrected and then assigned to the corresponding sample), discard low-quality reads and reads with uncalled sites and separate the reads based on their individual barcode information.

The genome project of one species of *Xiphophorus*, *X. maculatus*, has recently been completed (http://www.ncbi.nlm.nih.gov/genome/10764) (Schartl *et al.* in revision); however, here we used a *de novo* assembly approach implementing a two-step method to identify interspecies SNPs. First, within-species RAD tag loci were built using ustacks (Hohenlohe *et al.* 2010) in the Stacks pipeline where the minimum stack size was specified as 3 (-m 3), that is, minimum depth of coverage required to create a stack. Stacks were merged into individual loci by allowing a maximum of one mismatch between stacks (-m 1). Potentially incorrectly merged loci and highly repetitive stacks (lumberjack stacks) were excluded in the downstream analyses using the parameters -d and –r. We masked within-individual polymorphisms using the fixed SNP calling mode (-T fixed). In the second step, we used hstacks, from the Stacks pipeline, to identify interspecies SNPs through clustering and comparing potentially homologous loci from each individual data set. Mismatches among the RAD loci in each individual data set were identified using a maximum mismatch number of either five or eight.

SNP matrix preparation

We developed in-house Perl scripts (S. Fan) to parse the results of hstacks for both the species-level and individual-based matrices. The scripts first discarded all clusters containing sequences from the same individual as these sequences may represent duplicated regions in the genome. For the species-based analysis, intraspecies SNPs were merged using the standard degenerate code. The latter step was implemented as here we were particularly interested in identifying polymorphisms among species. The script also discarded those SNPs that were represented in fewer taxa than our specified minimum number of taxa. For the SNP matrix construction, we specified the minimum number of taxa as 15, 20 and 25 (corresponding to 52–86% of the samples). These minimum numbers of taxa were considered appropriate tests given the stochastic factors in genome shearing, PCR amplification and Illumina sequencing. These minimum numbers of taxa represent different potentially informative thresholds in the number of species included in the constructed stacks. A decrease in the minimum number of taxa is expected to increase the amount of missing data in the SNP matrix. We evaluated the impact of increasing the amount of missing data on our phylogenetic analyses in further analyses. Specifically we evaluated the impact of missing data by comparing supporting values and tree topologies that were based on matrices with different missing data thresholds.

For conducting an individual-based phylogenetic analysis, we constructed a SNP matrix with 149 362 SNP sites based on the hstacks result with maximum eight mismatches among potentially homologous regions in different individuals. The individual-based SNP script extracts the nucleotides at the reported SNP site from the hstacks output and does not merge polymorphisms as done in the species-level analysis. Two individuals (*X. gordoni* 1 and *X. pygmaeus* 5) were excluded for further analyses as these samples consisted of only missing data. A total of 139 individuals were used for further analyses. To minimize any effects of missing data in our phylogenetic analyses, we specified the minimum number of taxa as 100 in our individual-based analyses (corresponding to 71% of samples).

Phylogenetic analysis

To reconstruct the phylogenetic relationships among our study species, analyses were performed on all interspecies SNP matrices as outlined above (i.e. with the different parameter combinations) and also on the individual-based SNP matrix. Phylogenetic analyses were conducted using maximum likelihood (ML) and Bayesian inference and maximum parsimony on all interspecies SNP matrices. The individual SNP matrix-based analysis was conducted using ML. All ML analyses were conducted using the online version of RaxML (Stamatakis *et al.* 2008) with the general time-reversible (GTR) and gamma (Γ) (Yang & Kumar 1996) model of sequence evolution. The topology was evaluated with 100 'rapid-bootstrap' replicates (Stamatakis 2006). We conducted Bayesian analyses using BEAST v1.7.1 (Drummond & Rambaut 2007). For this analysis, we also implemented the GTR+gamma model of sequence evolution. To construct the phylogeny, the lognormal relaxed-clock model was used to allow rate variation among branches without a prior assumption (Drummond *et al.* 2006). For comparison, we also repeated the analysis with an estimated global molecular clock for all branches. In both analyses, a Yule speciation process (pure birth process) was implemented as the tree prior and 1×10^6 steps were used in the Markov chain Monte Carlo (MCMC) iterations with sampling every 1000 iterations. Convergence of the chain was confirmed by checking the effective sample size of the results using the program Tracer v1.5 (http://beast.bio.ed.ac.uk/Tracer). We tested whether the relaxed molecular clock improved the fit of models to the data compared with the global molecular clock by comparing the Bayes factors (Kass & Raftery 1995) of the likelihood of the results in Tracer. A single 'maximum clade credibility' tree, which is analogous to the majority-rule consensus tree estimation in ML analysis in PAUP, was generated using TreeAnnotator v1.7.1 (http://beast.bio.ed.ac.uk/TreeAnnotator) software using the sampling trees generated in BEAST. Maximum parsimony (MP) analyses were performed using PAUP* 4.0b10 (Swofford 2003) using heuristic searches with 10 random-addition-sequence replicates using the tree bisection-reconnection (TBR) branch swapping option. Bootstrap support values were calculated with 2000 replicates.

Divergence times were estimated using BEAST with a MCMC chain length of 2×10^7. Divergence estimations were obtained using lognormal relaxed-clock with the mean of the fastest and slowest rates of known calibrated molecular clocks in teleost fish (from 0.044 to 0.004 changes/site/Myr (Schories *et al.* 2009)). The results were parsed with Tracer.

Mapping RAD data to the X. maculatus genome

The raw reads of each *Xiphophorus* specimen, processed by process_radtags in the stacks pipeline, were mapped to the *X. maculatus* genome (GCA_000241075.1) using BWA v0.6.2 (Li & Durbin 2009). We implemented the default parameters in this program that tolerates up to four mismatches and one gap when aligning reads to the genome. The mapping results were processed using Samtools v 0.1.18 (Li *et al.* 2009) where the individual mapping results were merged into species-level results and ambiguously mapped reads were excluded. We implemented a minimum mapping quality score of >20 (−q 20).

To further investigate the hypothesis of hybridization in this genus, the genomic architecture of different potential parental lineages and the putative hybrid species were compared. In this comparison, species thought to be likely parental lineages, in addition to a selection of species from the different major clades, were compared with the hybrid species (i.e. the putative hybrid species *X. clemenciae* and *X. monticolus* were compared with the southern swordtail fishes *X. mixei*, *X. kallmani*, *X. hellerii*, the northern swordtails *X. montezumae* and *X. birchmani* and the southern platyfishes *X. maculatus, X. andersi and X. milleri*). The genomes of parental species are expected to harbour the highest number of similar regions with the hybrid species.

We calculated the similarity (%) of mapped regions between hybrid and potential parental species in a pairwise manner (e.g. *X. clemenciae* and *X. mixei, X. clemenciae* and *X. kallmani)* (calculated as: the number of mutations/length of the mapped region × 100). This comparison was based on mpileup files generated by Samtools with parameter C50 to downgrade mapping quality for reads containing excessive mismatches. In total, eight mpileup files were generated for each comparison. We implemented conservative parameters in this analysis to minimize the impact of any sequencing error. We excluded regions with <3× coverage and SNPs with <3× coverage. In addition, for the comparison between species pairs, we only included regions that overlapped among all eight comparisons. We note that this method necessarily results in a focus on regions that are conserved between species and the reference genome species (*X. maculatus*).

Ancestral state of the sword

The ancestral state of the sword was estimated using the ML and maximum parsimony methods using MESQUITE v2.72 (Maddison & Maddison 2010). A range of different sword traits are thought to influence mate choice in *Xiphophorus* fish (e.g. Trainor & Basolo 2006). Accordingly, following the study by Kang *et al.* (2013), we applied three different sword traits to our reconstruction of the sword: extension of the caudal fin, sword colouration and black pigmentation of the ventral margin (Meyer *et al.* 1994; Wiens & Morris 1996; Meyer 1997; Kang *et al.* 2013). We assigned the sword characters into three trait states in one analysis. First, we assigned species with a coloured sword to 'coloured sword', second we assigned species that are intermediate for any of the sword traits, that is, that are polymorphic for sword colour or ventral black margin or have a sword-like but smaller protrusion, to 'intermediate' and finally species with no evidence of any sword traits to 'no sword'. As a measure of statistical support for the ML ancestral state reconstruction analysis, we used the default likelihood threshold decision implemented in Mesquite ($T = 2$).

Results

Restriction site-associated DNA libraries were generated by individually barcoding and sequencing DNA from an average of five individuals of all 26 *Xiphophorus* species and three outgroups. Sequences were produced from sites throughout the genome specific to both *Msp*I and *Pst*I-HF restriction sites. We generated a total of 294 million raw reads from four Illumina GAIIx lanes and one Illumina HiSeq2000 lane (Table S1, Supporting information). After quality control, we included 237 million raw reads in our further analyses (Table S1, Supporting information). The average number of reads per individual in the data set was 87 290 ± 16 563 and ranged from 30 892 to 4 192 236 (Fig. S1, Supporting information). The average coverage achieved per individual per loci was 15x (Fig. S2, Supporting information). Due to the low coverage achieved for *X. evelynae*, individual samples for this species were combined for further analyses.

SNP matrices

As expected intuitively, SNP matrix size, that is, length of the alignment in bp, increases linearly with the permitted maximum number of mismatches in the homologous regions among different species (Fig. S3, Supporting information). In addition, a decrease in the minimum number of taxa represented at each locus produced a concomitant increase in the size of the SNP matrix. Under our intraspecies combined analysis, we generated a total of six SNP matrices for the phylogenetic analyses (using three different thresholds for the minimum number of species specified—15, 20 and 25, and two mismatch thresholds—5 and 8) with a range in total sequence length of 15 632 bp to 66 983 bp. The proportion of missing data per species in these matrices ranged from 1 to 78% (Table S2, Supporting information). One of the outgroups used in our study (*Heterandria formosa*) consistently had the highest proportion of missing data. This may be caused by a large molecular divergence between *H. formosa* and the rest of the species analysed, meaning the maximum number of mismatches allowed here was exceeded in comparisons among homologous regions between this outgroup species, and the rest of the study species. The latter result indicates that the RAD method is currently more appropriate for handling phylogenetic research on recently derived species (see also Rubin *et al.* 2012). Specifically, divergences among the genera studied here likely represent an upper bound on species divergence levels for utilizing RAD methodology.

Phylogenetic analyses

Using ML, Bayesian and MP analyses, the RAD SNP data set enabled the production of a highly supported tree topology, where the topology was very similar regardless of the analysis method used, the minimum number of species represented at each SNP site or the number of allowed mismatches between potentially homologous loci (Fig. 1). We found only minimal differences in supporting values at a few of the more derived branches (e.g. Fig. S4, Supporting information), and we also note there is a difference in topology in the derived northern platyfish clade (at two of six tested ML parameters) and the *X. clemenciae* clade (at one of six tested ML parameters) (Fig. S4, Supporting information). The amount of missing data had no significant impact on

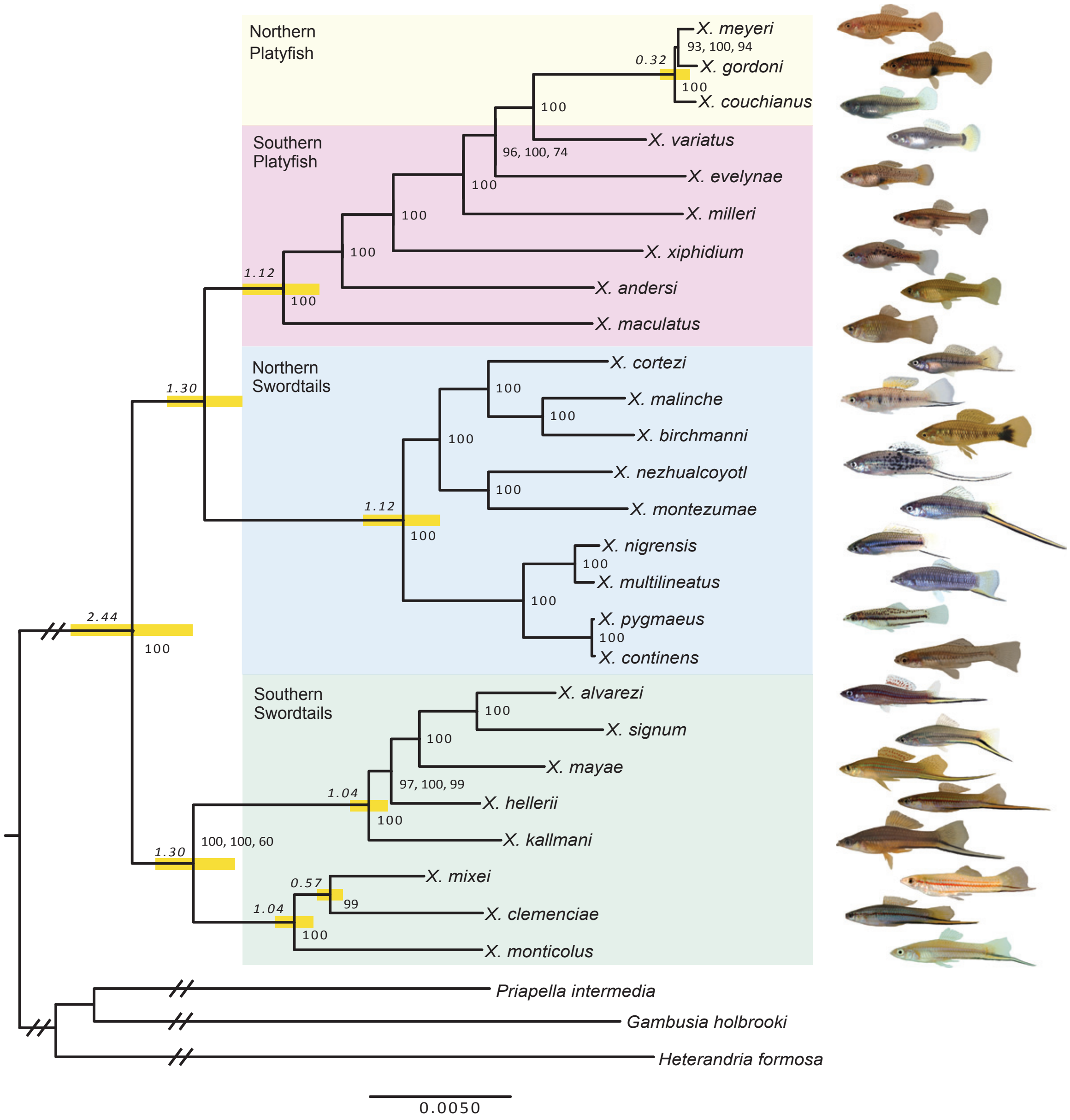


Fig. 1 Phylogenetic tree (ML, Bayesian and MP) of all *Xiphophorus* species based on the largest SNP matrix analysed (i.e. where SNPs were included if present in a minimum of 15 species and with a maximum of eight mismatches). When support values are 100 for all analyses, the value is listed once, otherwise support values are listed in the order ML, Bayesian, MP. Divergence time estimates—in million years—were calculated using the fastest and slowest rates of known calibrated molecular clocks in teleost fish [from 0.044 to 0.004 changes/site/Myr (Schories *et al.* 2009)] in BEAST. Mean divergence times are indicated in italics, and the upper and lower 95% confidence intervals are shown using a yellow bar.

the placement and branch support for *X. evelynae*, a species with a high proportion of missing data. Importantly, the consistency between our parametric and parsimony analyses suggests that misleading topologies and biases recently highlighted to be associated with missing data in likelihood-based analyses (Simmons 2012a,b see also Thomson & Shaffer 2010; Wiens & Morrill 2011) are unlikely to have had an impact on our results. We note that similar to Wagner *et al.* (2012), we obtained the largest SNP matrix using the lowest threshold tested here (minimum of 15 species at each SNP site), with a maximum of eight mismatches in the homologous regions among different species, and it provided the highest bootstrap support (Fig. 1). Therefore, in line with what Wagner *et al.* (2012) found, we find that with a greater number of SNPs included in the data matrix (Fig. S3, Supporting information), we have the greatest number of highly supported nodes (with the lowest bootstrap support being 93%). In our Bayesian analyses, all nodes are highly supported using the different SNP matrices. Using the 2ln Bayes > 10 criterion (Kass & Raftery 1995), phylogenies constructed with the lognormal relaxed-clock model (Drummond *et al.* 2006) were found to be significantly more reliable than those constructed using the global molecular clock. In our individual-based ML analyses, we find that all intraspecies individuals are monophyletic, the groupings are highly supported (data not shown) and interspecies clusters are identical to the results shown in Fig. 1.

Phylogenetic groups and divergence times

The data set resolved three major clades of *Xiphophorus* (southern swordtails, northern swordtails and the platyfish—where the northern platyfish form a clade, whereas the southern platies are not monophyletic within the platyfish clade) (Fig. 1, see also Fig. S4, Supporting information). We find the northern platyfish to be the most derived species and form part of a single monophyletic grouping with the southern platyfish (Fig. 1). These results, alongside the divergence time analyses using BEAST, suggest a more ancient divergence of the southern and northern swordtails and a more recent origin of the northern platyfish (Fig. 1, Table S3, Supporting information). However, we note that the absence of suitable fossil calibration time points limits the reliability of our divergence time estimates.

The monophyletic grouping of the southern swordtails, including two of the recently described species (*X. monticolus* and *X. mixei*), provides the first highly supported estimate of the phylogenetic relationships among these species. This group of southern swordtails includes two species hypothesized to be of hybrid origin (*X. clemenciae* and *X. monticolus*) due to discordance between mtDNA- and nuclear-based trees (Meyer *et al.* 2006; Kang *et al.* 2013). We note that previous work suggests that *X. maculatus* is the most likely maternal lineage of both putative hybrid species (Meyer *et al.* 2006; Kang *et al.* 2013; see also Jones *et al.* 2012). Here, we also provide the first strongly supported phylogenetic estimate of the evolutionary relationships within the northern swordtail clade (Fig. 1).

Divergence between swordtails and platyfish

Our ancestral state reconstruction analyses suggest that the most likely state of the sword in the ancestral lineage of *Xiphophorus* is 'sworded' (Fig. 2). However, we find that the ancestral state remains unresolved under maximum parsimony. Under ML, the *Xiphophorus* ancestral lineage is likely to have had a coloured sword with a ventral black margin—two traits that have also been shown to be preferred by females in some *Xiphophorus* species (e.g. Trainor & Basolo 2006).

Nuclear- vs. mtDNA-based phylogenies

The nuclear DNA-based phylogeny estimated here using RAD data shows some incongruences with previously estimated mtDNA-based phylogenetic trees, where two species, *X. clemenciae* and *X. monticolus*, show different phylogenetic positions between the analyses based on the different molecular markers (see also Meyer *et al.* 1994, 2006; Kang *et al.* 2013). In the nuclear DNA-based tree, both species are grouped with the southern swordtails (Fig. 1), whereas in mtDNA-based trees, both species are grouped with the southern platyfish (Meyer *et al.* 2006; Kang *et al.* 2013). Such divergent placements can result from three main evolutionary scenarios including hybridization events with backcrossing, introgressive hybridization which in concert with different selective forces acting on the mitochondria vs. the nuclear genome could result in such differences (e.g. Ballard & Whitlock 2004; see also Fig. 3) and incomplete lineage sorting. Interestingly, the strongly supported monophyletic grouping of the *X. clemenciae* clade vs. the rest of the southern swordtails suggests that if these species are indeed of hybrid origin, the *X. mixei* lineage is likely to be the paternal lineage of the hybrid species.

The nature of the RAD data generated here enables a genome-wide comparison of the genetic architecture of the discordantly placed species and other closely related species likely to have played a role in their evolution via hybridization. We note that mtDNA-based comparisons are limited here due to the nuclear genome specificity of the RAD enzymes. Genomic regions that fit our stringent criteria were included in our hybrid–potential parental lineage comparisons and included 1102 regions for comparisons with *X. clemenciae* and 968 regions for comparisons with *X. monticolus*. We show that when mapped to the *X. maculatus* genome scaffolds, *X. mixei* is most similar to both hybrid species compared with other southern swordtails, northern swordtails and southern platyfish (Fig. 4). There does not appear to be a large difference in similarity between regions, although scaffolds 1 and 2 are more similar between *X. mixei* and the hybrid species (see Fig. 4 for *X. clemenciae* example). We note that these similarity estimates are likely to be underestimated as highly diverged regions (for example regions with more than four mismatches, as limited by the mapping program) in the swordtail fish genomes will not map to the *X. maculatus* reference genome. Interestingly, *X. andersi* was found to be the most similar southern platyfish to both putative hybrid species, although all southern

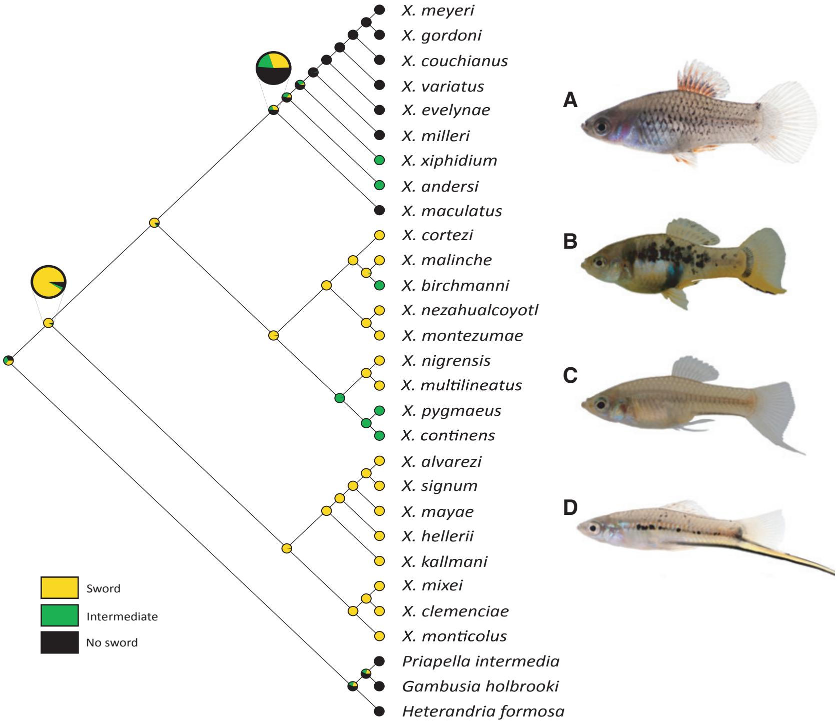


Fig. 2 Ancestral state reconstruction of the sexually selected sword trait under ML including proportional likelihood estimations for the alternative states. Sword characters are assigned to three trait states. First, species with a coloured sword were assigned to 'coloured sword', second species that are intermediate for any of the sword traits, that is, that are polymorphic for sword colour or ventral black margin or have a sword-like but smaller protrusion, were assigned to 'intermediate' and finally species with no evidence of any sword traits to 'no sword'. Examples of the different states are shown in the fish pictures [A, B, C, D, where A is an example of a species with no sword or colour (*Xiphophorus maculatus*); B is an example of a species with a small coloured protrusion (*Xiphophorus xiphidium*); C is an example of a species with a sword-like colourless protrusion (*Xiphophorus andersi*); D is an example of a species with a coloured sword (*Xiphophorus hellerii*)]. The two enlarged ancestral state estimations under ML are as follows: *Xiphophorus* ancestor is 0.906 'sworded', 0.031 'intermediate' and 0.063 for 'no sword'. These differences are significant at the default likelihood threshold in Mesquite (T = 2). The platyfish ancestral state is 0.515 'sworded', 0.188 'intermediate' and 0.297 'no sword'—this is not significant at the default likelihood threshold in Mesquite (T = 2). The ancestral state of the sword is unresolved using MP (data not shown).

platyfish show a relatively similar level of difference to the discordantly placed species (Fig. 4).

Discussion

Here, we resolve the evolutionary relationships among all species of *Xiphophorus* fish using thousands of base pairs of genome-wide SNP data. We provide support for three major clades (southern swordtails, northern swordtails and platyfish), and within the platyfish, a more recently derived northern platyfish clade. In contrast to previous estimates of the phylogenetic relationships among these fish where support and/or species information was lacking, we find strong support for the most basal split between the southern swordtails and the clade including both platyfish and the northern swordtails. Furthermore, we recover two distinct monophyletic groupings within both the southern swordtails and the northern swordtails.

We show that including a greater number of base pairs in the SNP matrix data set results in the highest support for all nodes. Importantly, we find that despite recently outlined possible caveats with missing data, which are particularly relevant to large NGS data sets (Simmons 2012a,b), here congruence between ML, Bayesian and MP analyses suggests our results are robust. However, below we outline cautions associated with phylogenomic analyses using RAD data—a currently young field. We provide further support for incongruence between the nuclear vs. mtDNA phylogenies of this genus which likely reflects the contribution of hybridization to the evolution of two species of *Xiphophorus* (*X. clemenciae* and *X. monticolus*). By reconstructing the evolutionary history of the sexually selected sword trait, our results suggest that coloured swords with a ventral black margin may be the ancestral state in the basal lineage of the genus *Xiphophorus*.

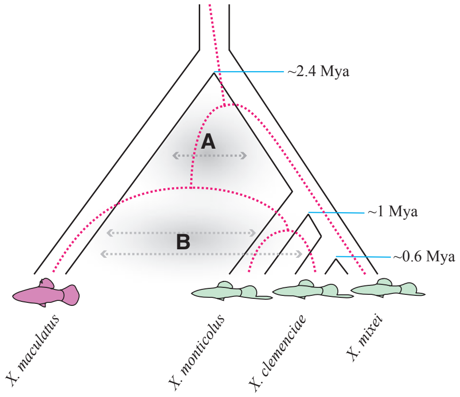
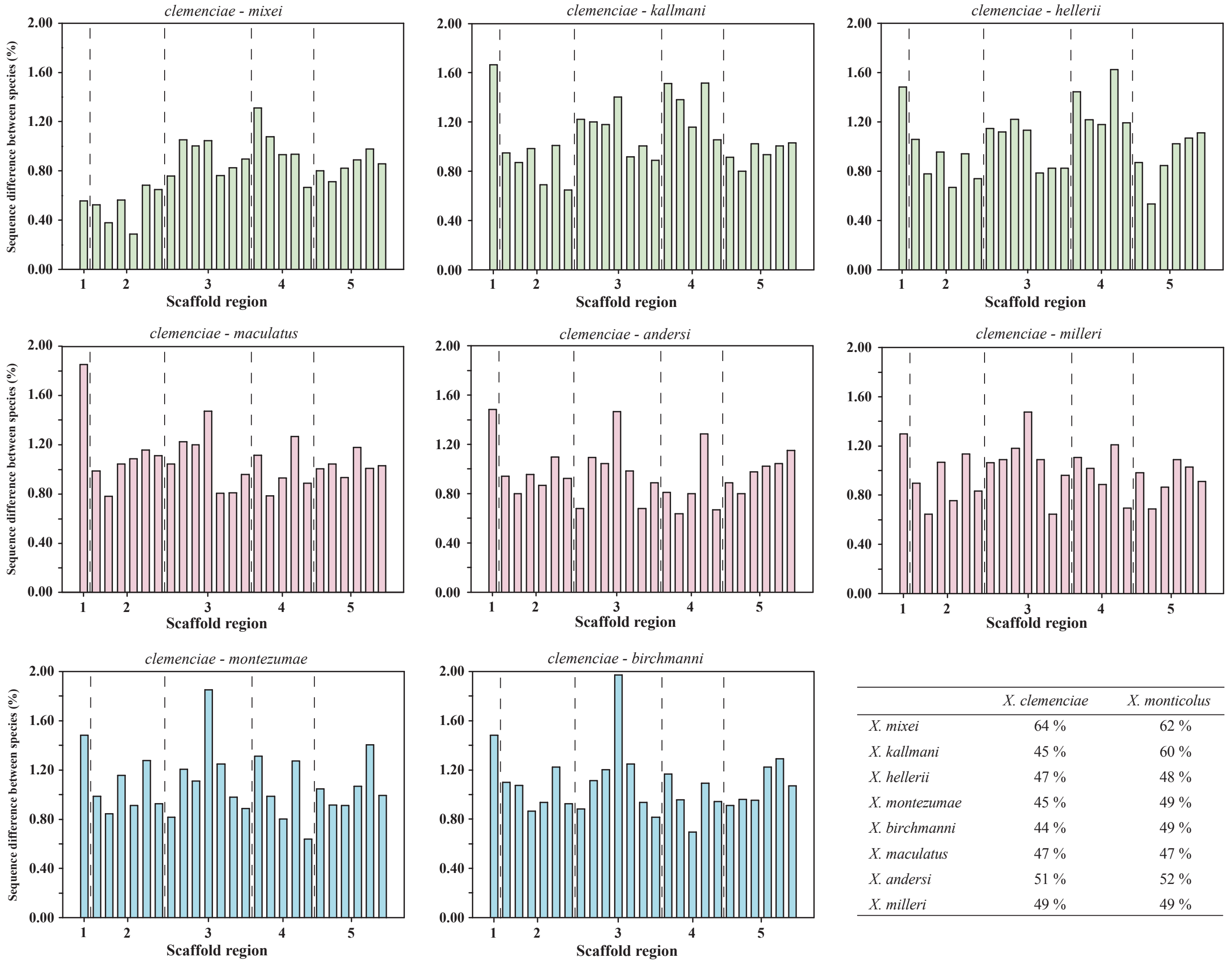


Fig. 3 Schematic diagram of phylogenetic discordance and evolutionary history of *Xiphophorus* fish. The black outline illustrates the RAD nuclear-based phylogeny, and the pink dashed line represents the mtDNA gene tree (mtDNA tree adapted from Kang *et al.* 2013). The figure represents possible hybridization and introgression scenarios in the evolutionary history of these fish leading to current discordance between molecular markers. Two hybridization scenarios can be hypothesized: hybridization (grey shading) between the southern swordtail and southern platyfish lineages (A); introgressive hybridization (grey shading) between *X. maculatus* and both *X. clemenciae* and *X. monticolus* resulting in the platyfish-like mtDNA haplotype introgressing into these species (B). Times (Mya) represent the estimated divergence times between species lineages.

Resolving the evolutionary history of swordtails and platyfish

The phylogenetic relationships among *Xiphophorus* fish has received much attention; however, different topologies have been recovered based on different sets of markers (Rosen 1960; Rosen & Kallman 1969; Rosen 1979; Rauchenberger *et al.* 1990; Basolo 1991; Meyer *et al.* 1994; Lockhart *et al.* 1995; Meyer *et al.* 2006; Kang *et al.* 2012). Using genetic markers throughout the genome to determine such evolutionary relationships allows a greater understanding of the evolutionary history of this group of fish. In a broad phylogeographic context, our molecular phylogeny provides support for the South to Central America radiation of Poeciliid fishes and the subsequent divergence of *Xiphophorus* species and populations (Hrbek *et al.* 2007). The basal split between the ancestor of the southern swordtails and the rest of the *Xiphophorus* radiation, and the younger origin of the platyfish clade relative to the southern swordtails (Fig. 1, Table S3, Supporting information), is in line with the known geographic distribution of these fishes (Kallman & Kazianis 2006). Specifically, the southern swordtails occur in distinct regions throughout the southernmost distribution of the *Xiphophorus* genus—Belize and Honduras north-west to Veracruz in Mexico—apart from *X. hellerii* which occurs throughout this entire range (Kallman & Kazianis 2006). Species in the *X. clemenciae* clade (*X. monticolus*, *X. clemenciae*, *X. mixei*), found to be one of the evolutionarily older groups of swordtail fish, interestingly have a very restricted distribution occurring only in the uplands of the Rio Coatzacoalcos Basin in the Isthmus of Tehuantepec (Kallman & Kazianis 2006). Thus, the current distribution of this group may be limited to specific upland refugial habitats. *Xiphophorus maculatus*, the most basal southern platyfish (Fig. 1), has a similarly broad distribution to *X. hellerii*. The majority of the southern platyfish occur in distinct regions neighbouring and then progressively north-west of the southern swordtails (Kallman & Kazianis 2006). However, the phylogenetic placement of *X. xiphidium* does not fit its known geographic distribution. Rather, this species has the most northern distribution of the southern platyfish and would therefore be expected to be phylogenetically placed between *X. variatus* and the northern platyfish (Fig. 1). The northern swordtails co-occur further to the north-west from most southern platies (although the southern platyfish, *X. variatus*, has a broad distribution that overlaps with the northern swordtails) (Kallman & Kazianis 2006). Fishes in the most recently derived clade, the northern platyfish, occur in distinct regions in the far northern part of the *Xiphophorus* distribution (Fig. 1, Kallman & Kazianis 2006).

With this large RAD determined SNP data set, we are able to resolve all evolutionary relationships of these fishes that have previously been contentious; however, we note current cautions with this data type below. Two distinct northern swordtail clades (the *X. pygmaeus* and *X. montezumae* clades), and the most recently derived nodes therein, are fully supported under all parameters tested here (Figs 1 and S4, Supporting information). The recent split between the *X. continens, X. pygmaeus* clade, and the *X. multilineatus, X. nigrensis* clade, is in line with their known ranges which include separate tributary rivers (*X. continens*: Rio Ojo Frio, *X. pygmaeus*: Rio Huchihuayan and Axtla, *X. multilineatus*: Rio Coy, *X. nigrensis*: Rio Choy) that flow into the larger river, Rio Tampoan, at different points along its course. The clear resolution achieved between *X. birchmanni* and *X. malinche* in particular is likely due to these species samples being originally obtained from sites outside of the current hybrid zone known for these species (e.g. Culumber *et al.* 2011). In addition, our data provide strong support for *X. maculatus* as the most basal southern platyfish, and *X. andersi* as the next most basal southern platyfish, a species that has been problematic to place phylogenetically. Together, these species distributions and new genome-wide molecular data provide a basis for further investigating genomic and geographic patterns of species



	X. clemenciae	*X. monticolus*
X. mixei	64 %	62 %
X. kallmani	45 %	60 %
X. hellerii	47 %	48 %
X. montezumae	45 %	49 %
X. birchmanni	44 %	49 %
X. maculatus	47 %	47 %
X. andersi	51 %	52 %
X. milleri	49 %	49 %

Fig. 4 Genomic regions shared between putative hybrid species and other *Xiphophrous* lineages (including potential parental lineages). Here, the putative hybrid species *X. clemenciae* is used as an example. Graphs show pairwise differences (percentages) between the putative hybrid species (*X. clemenciae* shown here) and putative parental lineages and representatives of the major *Xiphophorus* clades. Bars represent average differences (%) between species of mappable reads along each scaffold (i.e. where each bar is an average of ~50 regions of mappable reads in a sliding window approach). Colours designate the clade of each species compared here (see also Figure 1) and dotted vertical lines differentiate scaffold regions. The table indicates the percentage of regions that have the highest similarity score between species pairs (i.e. the table shows the percentage of cases where each specific species pair was the most similar of all eight pairwise comparisons).

diversification and also model-based tests of hybrid speciation (e.g. see Nice *et al.* 2013).

Phylogenetic estimations using RAD data

Here, we have inferred highly supported estimates of the evolutionary relationships among all species of the genus *Xiphophorus;* however, as suggested by Rubin *et al.* 2012 and Wagner *et al.* 2012, these results should be viewed with caution and examined further in future analyses. Our analyses here, which combine thousands of orthologous SNP markers throughout the genome of each species, do not account for any gene or region-specific selective forces acting on individual loci. However, the availability of the *X. maculatus* reference genome (http://www.ensembl.org/Xiphophorus_maculatus/Info/Index) and linkage information (Schartl *et al.*, submitted) will, in the future, likely enable the use of analysis methods that accommodate and investigate gene tree heterogeneity and infer species trees, such as BEST (Edwards *et al.* 2007; Liu 2008) and *BEAST (Heled & Drummond 2010) and BUCKy (Ané *et al.* 2007) see also Chung & Ané (2011), and more accurate model selection procedures. Analyses based on data concatenation may not reflect the true species tree (Kubatoko & Degnan 2007) and can sometimes produce unusually high levels of bootstrap support (Gadagkar *et al.* 2005; Seo 2008; Kumar *et al.* 2012). Partitioning data by gene or codon site could help to overcome biases caused by simplifications of assumptions across the genome (Rannala & Yang 2008). Further methods for partitioning NGS data are needed and, as is the case for *Xiphophorus*, will be aided by linkage and gene annotation information.

In addition, likelihood analyses can provide strong support for incorrect topologies when analyses are based on data sets with nonrandom distributions of missing data, as is the case with large phylogenetic data sets (Simmons 2012a,b a, b). In recent analyses using hypothetical examples, Simmons (2012a,b) showed such biases can occur even without mutation rate heterogeneity among characters and with a well-fitting model and when applied to simple data matrices. Using parsimony-based analyses rather than relying only on parametric methods is one approach recommended for testing whether likelihood-based results are likely to be artifactual (Simmons 2012a). Specifically, Simmons (2012a) suggests limiting phylogenetic inferences to branches that are supported in the parsimony strict consensus and investigating further any discrepancies between parsimony and parametric analyses. Here, we find no discrepancies in the results between different types of phylogenetic analysis suggesting that in this case, likelihood analyses are not adversely affected by missing data.

Evolution of a sexually selected trait

Previous estimations of the *Xiphophorus* phylogeny have raised two conflicting scenarios for the evolution of the sexually selected exaggerated male trait—the sword (e.g. Basolo 1995a,b; Meyer 1997). Traditionally, swordless platyfish have been estimated as basal and the swordtails (with swords) more derived (Rosen 1979; Rauchenberger *et al.* 1990), providing evidence for the trait having arisen in line with the pre-exiting bias hypothesis because it could be shown that 'already'—assuming their basal placement—platyfish females prefer swords and, thereby, might have driven the subsequent evolution of swords in platyfish.

Here, using ML analyses, we provide support for the alternative hypothesis that the sword, including its colouration, may have evolved in the immediate ancestral lineage of the genus, therefore also bringing into question whether the pre-existing bias in female preference accounts for the evolution of the sword trait (see also Kang *et al.* 2013; Fig. 2). Rather, the different characteristics, including basic presence, colour and also dark pigmented dorsoventral margins, have been reduced and lost in the evolutionary history of the genus. However, a pre-existing bias in female preference for the sword trait cannot be ruled out because the ancestral state of the outgroups remains ambiguous (Fig. 2). In addition, under MP, the ancestral lineage cannot be assigned to either a sworded or nonsworded state. Further, although females in a sister species of one outgroup, *Priapella olmecae*, are known to prefer sworded males over their own nonsworded males (Basolo 1998), the preferences of other closely related species remain untested. Identifying the driving force behind the evolution of the sword trait would benefit from implementing a greater array of female preference studies. We also note that the limitations in conducting phylogenetic analyses with RAD data, as detailed above, mean that the evolutionary history of the sword trait should also be considered with caution.

Nuclear- vs. mtDNA-based phylogenies

The high phylogenetic resolution achieved in this study using a large set of nuclear markers throughout the genome allows an informative comparison with previous mtDNA estimates of the evolutionary relationships among this group of fish. We find discordance between the RAD nuclear-based phylogeny estimated here (Fig. 1) and previously estimated mtDNA-based phylogenies (Meyer *et al.* 1994, 2006; Kang *et al.* 2013; see also Fig. 3). Specifically, in the RAD-marker-based tree, both species group with the southern swordtail clade—a result that is expected based on previous DNA studies and their phenotypic resemblance to other southern swordtails. In the mtDNA-based tree, these species grouped with the southern platyfish. Three main evolutionary scenarios, which are not necessarily mutually exclusive, are likely to give rise to such a phylogenetic pattern (see Fig. 3). First, *hybridization with backcrossing*, where as hypothesized by Meyer *et al.* (1994, 2006), the discordantly placed species may be good hybrid species resulting from hybridization between a southern swordtail and a southern platyfish (likely the maternal species) followed by repeated backcrossing with southern swordtail males. Incongruence between molecular phylogenies based on different marker types is a recognized signature of past hybridization (Arnold 1992; Avise 1994, 2000; Seehausen 2004; Arnold & Meyer 2006). Such differences between marker sets, supported here by a nuclear-based phylogeny, along side previously documented behavioural characteristics of the putative parental lineages and intermediate morphology of the putative hybrid *X. clemenciae* (Meyer *et al.* 2006), provide support for hybridization events giving rise to these species. In addition, the large degree of similarity throughout the mappable regions of the genome, between *X. mixei* and the putative hybrid species (Fig. 4), suggests that hybridization followed by backcrossing is a possible evolutionary scenario leading to such discordance.

A second related mechanism is *introgressive hybridization*, where even infrequent gene flow between different species in addition to the basic characteristic differences between mitochondrial and nuclear DNA, and the actions of selection or drift, can result in conflicts between mitochondrial and nuclear data (e.g. Shaw 2002; Sota 2002; Rognon & Guyomard 2003; Ballard & Whitlock 2004; Good *et al.* 2008; Barbanera *et al.* 2009; Pustovrh *et al.* 2011; Miller *et al.* 2012). Here, a putative scenario is that introgression occurred between the incongruently placed species lineages and a platyfish. At the same time, there may have been selection for the platyfish mtDNA haplotype, either direct or indirect, under the particular habitat conditions resulting in the platy-like haplotype introgressing into the genomes of *X. clemenciae* and *X. monticolus* and becoming fixed. We note that other demographic effects, rather than strict selection, can also promote such patterns (e.g. see Excoffier & Ray 2008). The small size of the mitochondrial genome, the lack of recombination in the genome and the maternal nature of inheritance mean that it likely represents a single molecular history, and in some cases the mtDNA from one taxon can completely replace that of another (i.e. mitochondrial DNA capture), without evidence of nuclear introgression or morphological signal (Ballard & Whitlock 2004; e.g. Bernatchez *et al.* 1995). Importantly, even without introgression, differences in divergence patterns between nuclear and mitochondrial DNA often come about because of the evolutionary properties of these DNA classes (Ballard & Whitlock 2004; Rheindt & Edwards 2011). Future investigations will be required to determine whether indeed there is selection for a platy-like mitochondrial genome in natural populations of the incongruently places species. Additionally, a mitochondrial genome-wide phylogenetic analysis of all *Xiphophorus* species may allow a high-resolution estimate of the evolution of this molecule among an entire genus. Further, with the completion of the *Xiphophorus* genome assembly and annotation, it will become possible to determine which genomic regions support the different evolutionary relationships and what their functions are.

Third, discordance between the nuclear and mitochondrial DNA-based trees may be due to incomplete lineage sorting. This possibility cannot be ruled out and will be resolved further through a sliding window-based analysis proposed above; however, there are two main reasons why this scenario is less likely. First, current geographic distributions of the species align well with the nuclear-based RAD phylogeny but similarly support the mtDNA-based phylogeny. In particular, overlapping distributions between the widely occurring platyfish, *X. maculatus*, and the more restricted distributions of the incongruent species suggest introgression likely accounts for the observed phylogenetic discordance rather than incomplete lineage sorting. Yet, we note that such introgression has not been reported so far, and our detailed phylogeographic analysis and sampling did not find a single instance of introgression between *X. maculatus* and *X. clemenciae* (Jones *et al.* 2012). Second, here we find that in ~63% of regions throughout the genome, *X. mixei* has the highest similarity with the incongruently placed species, compared with only about 50% with the platyfish (Fig. 4). This suggests that the RAD tree presented here represents a true species tree, while introgression with a platyfish, or the differential evolutionary history of the mitogenome, led to the discordance observed between the mitochondrial and nuclear genomes. Although mtDNA has a smaller effective size and therefore is expected to resolve in phylogenies faster, the evolutionary properties of this maternally nonrecombining molecule can differ from the true species history.

Under the scenario that *X. clemenciae* and *X. monticolus* arose via hybridization with backcrossing, our data suggest that the *X. mixei* lineage is likely to be the paternal lineage of both putative hybrid species (Fig. 1). Different to previous work (Meyer *et al.* 2006), the high support for the monophyly of the clemenciae clade recovered here strongly supports this inference. In addition, this putative paternal species overlaps in its current geographic distribution with both putative hybrid species (Kallman & Kazianis 2006), and gonopodium characters between all three species are similar (Jones *et al.* unpublished). Similar to the two hybrid species, *X. mixei* has a restricted distribution suggesting the hypothesis of a single origin for both species is likely, although this assumption is contingent on current distributions reflecting past distributions of the *X. mixei* lineage. Our clock estimates suggest that the clemenciae clade diverged one million years ago, meaning we can broadly infer that the incongruently placed species diverged at this time. Interestingly, *X. mixei* has a shorter sword than both hybrid species suggesting that selection for longer swords was possibly weaker in this lineage after hybridization occurred (see Fig. 2).

Additional behavioural studies are required to provide further insight into the mechanisms that promoted the possible hybrid origin of these two *Xiphophorus* species. The latter will also allow fruitful temporal comparisons with documented ongoing hybridization in the genus that may be examples of insipient speciation, such as between *X. malinche* and *X. birchmanni* (Culumber *et al.* 2011; Rosenthal & García de León 2011).

Acknowledgements

We thank L. Beck, A. Semmler and M. Weber for assistance with laboratory work and H. Recknagel for helpful comments on the study. We thank the *Xiphophorus* Genetic Stock Centre (San Marcos) for supplying samples of *X. mixei* and *X. monticolus* and Gil Rosenthal for supplying the *X. malinche* samples. We also thank our three anonymous reviewers for their comments, suggestions and insights that helped in improving the manuscript. This work was supported by grants of the Deutsche Forschungsgemeinschaft (DFG) (JCJ, AM), the Zukunftskolleg (JCJ), the University of Konstanz (AM) and the University of Würzburg (MS).

References

Ané C, Larget B, Baum DA, Smith SD, Rokas A (2007) Bayesian estimation of concordance among gene trees. *Molecular Biology and Evolution*, **24**, 412–426.

Arnold ML (1992) Natural hybridization as an evolutionary process. *Annual Review of Ecology and Systematics*, **23**, 237–261.

Arnold ML, Meyer A (2006) Natural hybridization in primates: one evolutionary mechanism. *Zoology*, **109**, 261–276.

Avise JC (1994) *Molecular Markers, Natural History and Evolution*. Chapman and Hall, New York.

Avise JC (2000) *Phylogeography – The History and Formation of Species*. Harvard University Press, Cambridge, Massachusetts.

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

Ballard JWO, Whitlock MC (2004) The incomplete natural history of mitochondria. *Molecular Ecology*, **13**, 729–744.

Barbanera F, Zuffi MAL, Guerrini M *et al.* (2009) Molecular phylogeography of the asp viper *Vipera aspis* (Linneaus, 1758) in Italy: evidence for introgressive hybridization and mitochondrial DNA capture. *Molecular Phylogenetics and Evolution*, **52**, 103–114.

Basolo AL (1990a) Female preference for male sword length in the green swordtail *Xiphophorus hellerii* (Pisces: Poeciliidae). *Animal Behaviour*, **40**, 332–338.

Basolo AL (1990b) Female preference predates the evolution of the sword in swordtail fish. *Science*, **250**, 808–810.

Basolo AL (1991) Male swords and female preferences: response. *Science*, **253**, 1426–1427.

Basolo AL (1995a) A further examination of a preexisting bias favoring a sword in the genus *Xiphophorus*. *Animal Behaviour*, **50**, 365–375.

Basolo AL (1995b) Phylogenetic evidence for the role of a preexisting bias in sexual selection. *Proceedings of the Royal Society of London B*, **259**, 307–311.

Basolo AL (1998) Evolutionary change in a receiver bias: a comparison of female preference functions. *Proceedings of the Royal Society of London B*, **265**, 2223–2228.

Bernatchez L, Glémet H, Wilson CC, Dazman RG (1995) Introgression and fixation of Arctic char (*Salvelinus alpinus*) mitochondrial genome in an allopatric population of brook trout (*Salvelinus fontinalis*). *Canadian Journal of Fisheries and Aquatic Science*, **52**, 179–185.

Catchen JM, Amore 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.

Chung Y, Ané C (2011) Comparing two Bayesian methods for gene tree/species tree reconstruction: simulations with incomplete lineage sorting and horizontal gene transfer. *Systematic Biology*, **60**, 261–275.

Culumber ZW, Fisher HS, Tobler M *et al.* (2011) Replicated hybrid zones of *Xiphophorus* swordtails along an elevational gradient. *Molecular Ecology*, **20**, 342–356.

Dasmahapatra KK, Walters JR, Briscoe AD *et al.* (2012) Butterfly genome reveals promiscuous exchange of mimicry adaptations among species. *Nature*, **487**, 94–98.

Drummond AJ, Rambaut A (2007) BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evolutionary Biology*, **7**, 214.

Drummond AJ, Ho SYW, Phillips MJ, Rambaut, (2006) Relaxed phylogenetics and dating with confidence. *PLoS Biology*, **4**, e88.

Edwards SV, Liu L, Pearl K (2007) High-resolution species trees without concatenation. *Proceedings of the National Academy of Science USA*, **104**, 5936–5941.

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

Excoffier L, Ray N (2008) Surfing during population expansions promotes genetic revolutions and structuration. *Trends in Ecology and Evolution*, **23**, 347–351.

Gadagkar SR, Rosenberg MS, Kumar S (2005) Inferring species phylogenies from multiple genes: concatenated sequence tree versus consensus gene tree. *Journal of Experimental Zoology Part B: Molecular Developmental Evolution*, **304**, 64–74.

Good JM, Hird S, Reid N *et al.* (2008) Ancient hybridization and mitochondrial capture between two species of chipmunks. *Molecular Ecology*, **17**, 1313–1327.

Heled J, Drummond AJ (2010) Bayesian inference of species trees from multilocus data. *Molecular Biology and Evolution*, **27**, 570–580.

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.

Hohenlohe PA, Amish SJ, Catchen JM, Allendorf FW, Luikart G (2011) Next-generation RAD sequencing identifies thousands of SNPs for assessing hybridization between rainbow and westslope cutthroat trout. *Molecular Ecology Resources*, **11**, 117–122.

Hohenlohe PA, Bassham S, Currey M, Cresko WA (2012) Extensive linkage disequilibrium and parallel adaptive divergence across threespine stickleback genomes. *Philosophical Transactions of the Royal Society B-Biological Sciences*, **367**, 395–408.

Hrbek T, Seckinger J, Meyer A (2007) A phylogenetic and biogeographic perspective on the evolution of poeciliid fishes. *Molecular Phylogenetics and Evolution*, **43**, 986–998.

Jones JC, Perez-Sato J-A, Meyer A (2012) A phylogeographic investigation of the hybrid origin of a species of swordtail fish from Mexico. *Molecular Ecology*, **21**, 2692–2712.

Kallman KD, Kazianis S (2006) The genus Xiphophorus in Mexico and Central America. *Zebrafish*, **3**, 271–285.

Kang JH, Schartl M, Meyer A (2013) Comprehensive phylogenetic analysis of all species of swordtails and platies (Pisces: Genus Xiphophorus) uncovers a hybrid origin of a swordtail fish, *Xiphophorus monticolus*, and demonstrates that the sexually selected sword originated in the ancestral lineage of the genus, but was lost again secondarily*BMC Genomics*, **13** (in press).

Kass RE, Raftery AE (1995) Bayes factors. *Journal of the American Statistical Association*, **90**, 773–795.

Kirkpatrick M, Ryan MJ (1991) The evolution of mating preferences and the paradox of the lek. *Nature*, **350**, 33–38.

Kubatoko LS, Degnan JH (2007) Inconsistency of phylogenetic estimates from concatenated data under coalescence. *Systematic Biology*, **56**, 17–24.

Kumar S, Filipski AJ, Battistuzzi FU, Kosakovsky Pond SL, Tamura K (2012) Statistics and truth in phylogenomics. *Molecular Biology and Evolution*, **29**, 457–472.

Lampert KP, Schmidt C, Fischer P *et al.* (2010) Determination of onset of sexual maturation and mating behaviour by melanocortin receptor 4 polymorphisms. *Current Biology*, **20**, 1729–1734.

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.

Liu L (2008) BEST: Bayesian estimation of species trees under the coalescent model. *Bioinformatics*, **24**, 2542–2543.

Lockhart PJ, Penny D, Meyer A (1995) Testing the phylogeny of swordtail fishes using split decomposition spectral analysis. *Journal of Molecular Evolution*, **41**, 666–674.

Maddison WP, Maddison DR (2010) Mesquite: a modular system for evolutionary analysis. Available at: http://mesquiteproject.org.

McCoy E, Syska N, Plath M, Schlupp I, Riesch R (2011) Mustached males in a tropical poeciliid fish: emerging female preference selects for a novel males trait. *Behavioural Ecology and Sociobiology*, **65**, 1437–1445.

Meyer A (1997) The evolution of sexually selected traits in male swordtail fishes (*Xiphophorus*: Poeciliidae). *Heredity*, **79**, 329–337.

Meyer A, Morrissey JM, Schartl M (1994) Recurrent origin of a sexually selected trait in *Xiphophorus* fishes inferred from molecular phylogeny. *Nature*, **386**, 539–542.

Meyer A, Salzburger W, Schartl M (2006) Hybrid origin of a swordtail species (Teleostei: *Xiphophorus clemenciae*) driven by sexual selection. *Molecular Ecology*, **15**, 721–730.

Miller W, Schuster SC, Welch AJ *et al.* (2012) Polar and brown bear genomes reveal ancient admixture and demographic footprints of past climate change. *Proceedings of the National Academy of Science USA*, **109**, E2382–E2390.

Nice CC, Gompert Z, Fordyce JA, Forister L, Lucas LK, Buerkle A (2013) Hybrid speciation and independent evolution in lineages of alpine butterflies. *Evolution* doi: 10.1111/evo.12019.

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 nonmodel species. *PLoS ONE*, **7**, e37135.

Pustovrh G, Susnik Bajec S, Snoj A (2011) Evolutionary relationship between marble trout of the northern and the southern Adriatic basin. *Molecular Phylogenetics and Evolution*, **59**, 761–766.

Quattro JM, Vrijenhoek RC (1989) Fitness differences among remnant populations of the endangered sonoran topminnow. *Science*, **248**, 976–978.

Quattro JM, Leberg PL, Douglas ME, Vrijenhoek RC (1996) Molecular evidence for a unique evolutionary lineage of endangered Sonoran desert fish (genus *Poeciliopsis*). *Conservation Biology*, **10**, 128–135.

Rannala B, Yang Z (2008) Phylogenetic inference using whole genomes. *Annual Review of Genomics and Human Genetics*, **9**, 217–231.

Rauchenberger M, Kallman KD, Morizot DC (1990) Monophyly and geography of the Rio Panuco Basin swordtails (genus *Xiphophorus*) with descriptions of four new species. *American Museum Noviates*, **2975**, 1–41.

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

Rheindt FE, Edwards S (2011) Genetic introgression: an integral but neglected component of speciation in birds. *The Auk*, **128**, 620–632.

Rognon X, Guyomard R (2003) Large extent of mitochondrial DNA transfer from *Oreochromis aureus* to *O. niloticus* in West Africa. *Molecular Ecology*, **12**, 435–445.

Rosen DE (1960) Middle-American poeciliid fishes of the genus *Xiphophorus*. *Bulletin of the Florida State Museum*, **5**, 57–242.

Rosen DE (1979) Fishes from the uplands and inter-montane basins of Guatemala: revisionary studies and comparative geography. *Bulletin of the American Museum of Natural History*, **162**, 268–375.

Rosen DE, Kallman KD (1969) A new fish of the genus *Xiphophorus* from Guatemala, with remarks on the taxonomy of endemic forms. *American Museum Noviates*, **2379**, 1–29.

Rosenthal GG, Evans CS (1998) Female preference for swords in *Xiphophorus hellerii* reflects a bias for large apparent size. *Proceedings of the National Academy of Science USA*, **95**, 4431–4436.

Rosenthal GG, García de León FJ (2011) Speciation and hybridization. In: *Ecology and Evolution of Poeciliid Fishes*(eds Schlupp I, Pilastro A & Evans J), pp. 109–119. University of Chicago Press, Chicago, Illinois.

Rosenthal GG, Wagner WE Jr, Ryan MJ (2002) Secondary reduction of preference for the sword ornament in the pygmy swordtail *Xiphophorus nigrensis* (Pisces: Poeciliidae). *Animal Behaviour*, **63**, 37–45.

Rubin BER, Ree RH, Moreau CS (2012) Inferring phylogenies from RAD sequence data. *PLoS ONE*, **7**, e33394.

Ryan MJ, Wagner WEJ (1987) Asymmetries in mating preferences between species: female swordtails prefer heterospecific males. *Science*, **236**, 595–597.

Schartl M (2004) Sex chromosome evolution in non-mammalian vertebrates. *Genetics and Development*, **14**, 634–641.

Schartl M (2008) Evolution of *Xmrk*: an oncogene, but also a speciation gene? *BioEssays*, **30**, 822–832.

Schories S, Meyer MK, Schartl M (2009) Description of *Poecilia* (*Acanthophacelus*) *obscura* n. sp., (Teleostei: Poeciliidae), a new guppy species from western Trinidad, with remarks on *P. wingei* and the status of the "Endler's guppy". *Zootaxa*, **2266**, 35–50.

Seehausen O (2004) Hybridization and adaptive radiation. *Trends in Ecology and Evolution*, **19**, 198–207.

Seo TK (2008) Calculating bootstrap probabilities of phylogeny using multilocus sequence data. *Molecular Biology and Evolution*, **25**, 960–971.

Shaw KL (2002) Conflict between nuclear and mitochondrial DNA phylogenies of a recent species radiation: what mtDNA reveals and conceals about modes of speciation in Hawaiian crickets. *Proceedings of the National Academy of Science USA*, **99**, 16122–16127.

Shen Y, Catchen J, Garcia T *et al.* (2012) Identification of transcriptome SNPs between *Xiphophorus* lines and species for assessing allele specific gene expression within F_1 interspecies hybrids. *Comparative Biochemistry and Physiology Part C: Toxicology and Pharmacology*, **155**, 102–108.

Simmons MP (2012a) Misleading results of likelihood-based phylogenetic analyses in the presence of missing data. *Cladistics*, **28**, 208–222.

Simmons MP (2012b) Radical instability and spurious branch support by likelihood when applied to matrices with non-random distribution of missing data. *Molecular Phylogeny and Evolution*, **62**, 472–484.

Sota T (2002) Radiation and reticulation: extensive introgressive hybridization in the carabid beetles *Ohomopterus* inferred from mitochondrial gene genealogy. *Population Ecology*, **44**, 145–156.

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

Stamatakis A, Hoover P, Rougemont J (2008) A rapid bootstrap algorithm for the RAxML Web Servers. *Systematic Biology*, **57**, 758–771.

Stöck M, Lampert KP, Möller D, Schlupp I, Schartl M (2010) Monophyletic origin of multiple clonal lineages in an asexual fish (*Poecilia formosa*). *Molecular Ecology*, **19**, 5204–5215.

Swofford D (2003) PAUP*. *Phylogenetic Analysis Using Parsimony (*and Other Methods)* Sinauer Associates, Sunderland, MA.

Thomson RC, Shaffer HB (2010) Sparse supermatrices for phylogenetic inference: taxonomy, alignment, rogue taxa, and the phylogeny of living turtles. *Systematic Biology*, **59**, 42–58.

Trainor BC, Basolo AL (2006) Location, location, location: stripe position effects on female sword preference. *Animal Behaviour*, **71**, 135–140.

Volff J-N, Schartl M (2001) Variability of genetic sex determination in poeciliid fishes. *Genetica*, **111**, 101–110.

Vrijenhoek RC, Douglas MC, Meffe GK (1985) Conservation genetics of endangered fish populations in Arizona. *Science*, **229**, 400–402.

Wagner C, Keller I, Wittwer S *et al.* (2012) Genome wide RAD sequencing data provides unprecedented resolution of species boundaries and relationships in the Lake Victoria cichlid adaptive radiation. *Molecular Ecology* **22**, 787–798.

Wiens JJ, Morrill MC (2011) Missing data in phylogenetic analysis: reconciling results from simulations and empirical data. *Systematic Biology*, **60**, 719–731.

Wiens JJ, Morris MR (1996) Character definitions, sexual selection, and the evolution of swordtails. *The American Naturalist*, **147**, 866–869.

Willing E-M, Bentzen P, Van Oosterhout C *et al.* (2010) Genome-wide single nucleotide polymorphisms reveal population history and adaptive divergence in wild guppies. *Molecular Ecology*, **19**, 968–984.

Yang Z, Kumar S (1996) Approximate methods for estimating the pattern of nucleotide substitution and the variation of substitution rates among sites. *Molecular Biology and Evolution*, **13**, 650–659.

J.C.J. is an evolutionary biologist interested in hybridization and species diversification. S.F. is a PhD student working on phylogenetic analyses using next-generation sequence data. P.F. is an evolutionary biologist interested in genome-wide analyses, population genetics and speciation. M.S. is the head of the Department of Physiological Chemistry at the Biocenter at the University of Würzburg and is interested in the molecular processes in the development of organisms and their malfunction in cancerogenesis. A.M. is the chair in evolutionary biology at the University of Konstanz and is interested in genomic, developmental and morphological aspects of speciation.

Data accessibility

Table 1 provides all sample data. All raw sequence data can be found at SRA065395. All SNP matrices used in the different phylogenetic analyses can be found in Dryad: doi:10.5061/dryad.728b4.

Supporting information

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

Fig. S1 Number of raw reads obtained per individual for each of the 29 species in this analysis.

Fig. S2 Coverage obtained per individual. Due to the low number of reads achieved for *X. evelynae*, individual samples for this species were combined for further analyses.

Fig. S3 SNP matrix size found when different numbers of maximum mismatches (5 or 8) and minimum number of species (15, 20 or 25) are implemented.

Fig. S4 Phylogenetic estimations using ML with different maximum mismatches and minimum number of species (the following order is shown at each node when estimations differed from bootstrap values of 100—m5, sp15; m5, sp20; m5, sp25; m8, sp15; m8, sp20; m8, sp25). *indicates a different topology at that node; in the clemenciae clade, X. monticolus and X. clemenciae were found to be the most derived species using one of six matrices. In the northern platyfish clade, X. couchianus and X. gordoni were found to be the most derived using two of six matrices. These estimations were less well supported than the topology found using the majority of the SNP matrices.

Table S1 Overview of RAD raw data and experimental setup. Individual raw read data, read data retained after quality control, library number (multiplexing design) and sequencing platform.

Table S2 Proportion of missing data at all parameters analysed in this study where maximum and minimum numbers are calculated for the ingroup taxa (see also Fig. S4).

Table S3 Estimated time of origin of the major *Xiphophorus* clades including mean and 95% confidence interval values (see also Methods).