Benefits and limitations of a new genome-based PCR-RFLP genotyping assay (GB-RFLP): a SNP-based detection method for identification of species in extremely young adaptive radiations

Claudius Kratochwil[1], Andreas Kautt[2], Sina Rometsch[2], and Axel Meyer[3]

[1]University of Helsinki
[2]University of Konstanz
[3] University of Konstanz

February 1, 2022

Abstract

High-throughput DNA sequencing technologies make it possible now to sequence entire genomes relatively easily. Complete genomic information obtained by whole genome resequencing (WGS) can aid in identifying and delineating species even if they are extremely young, cryptic or morphologically difficult to discern and closely related. Yet for taxonomic or conservation biology purposes WGS can remain cost-prohibitive, too time-consuming, and often constitute a "data overkill". Rapid and reliable identification of species (and populations) that is also cost-effective is possible based on species-specific markers that can be discovered by WGS. Based on WGS data we designed a PCR restriction fragment length polymorphism (PCR-RFLP) assay for 19 Neotropical Midas cichlid populations (Amphilophus cf. citrinellus), that includes all 13 described species of this species complex. Our work illustrates that identification of species and populations (i.e., fish from different lakes) can be greatly improved by designing genetic markers using available "high resolution" genomic information. Yet, our work also shows that even in the best-case scenario, when whole-genome resequencing information is available, unequivocal assignments remain challenging when species or populations diverged very recently, or gene flow persists. In summary, we provide a comprehensive workflow on how to design RFPL markers based on genome re-sequencing data, how to test and evaluate their reliability, and discuss the benefits and pitfalls of our approach.

1. Introduction

Historically morphological differences remain the basis for species identification, taxonomic keys, and effort in species delimitation. Yet, reliable classification of specimens can be complex due to many factors. For example, when species are morphologically extremely similar or when morphological characters are not expressed at a given life-history stage (e.g., juveniles). In the last decade, the increasing affordability of reduced-representation data (e.g., restriction-site-associated DNA sequencing or target enrichment) or whole genome re-sequencing has provided new possibilities to assign species not only based on morphological or meristic characters, but also on genomic information. In some instances, this has even greatly contributed to the discovery and description of new (i.e., previously cryptic) species (Fennessy et al., 2016; Nater et al., 2017). Genetic species assignment approaches are also promising to add novel tools to aid in conservation efforts of endangered species, but practical implementations often fail (Campbell et al., 2019; Piertney, 2016; Shafer et al., 2015). A major disadvantage of high-throughput sequencing techniques are the cost and time that is needed to generate libraries, sequence them, and to analyze the data. But, importantly, genomic data also allow for the identification of a suite of informative, diagnostic genetic markers for species or population assignment that can be genotyped using cheaper and faster methods (Shafer et al., 2015).

Among all genetic variants, single-nucleotide polymorphisms (SNPs) are clearly the most abundant (in the human population for example more than 95% of all genetic variants are SNPs (Auton et al., 2015)) and therefore powerful genetic markers for assigning populations or species. Over the past 30 years, many methods have been developed to cost-effectively genotype SNPs. One widely used fast method are PCR restriction fragment length polymorphism (PCR-RFLP) markers (McKeown, Robin, & Shaw, 2015; Ota, Fukushima, Kulski, & Inoko, 2007). Hereby, a particular DNA fragment is first amplified by PCR. The resulting amplicon is then digested using a restriction enzyme that cuts only one allele at a diagnostic SNP (resulting in two fragments) but not the other one (one fragment), due to an, ideally species-specific, polymorphism in the enzyme's recognition site. Homozygous individuals for either allele, as well as heterozygous individuals (three fragments), can be easily distinguished from each other by gel electrophoresis (see detailed description of the method in Ota et al., 2007). Therefore, PCR-RFLP is an excellent method that can be used for fast, cheap, and reliable genotyping of diagnostic markers.

Recently, we have sequenced 453 genomes of a very young species flock of Nicaraguan Midas cichlid fishes (*Amphilophus cf. citrinellus*) (Kautt et al., 2020). This species complex includes, so far, 13 described species (Torres-Dowdall & Meyer, in press). Two species (*A.s citrinellus* and *A. labiatus*) can be found in both Great Lakes Managua and Nicaragua (Barluenga, Stölting, Salzburger, Muschick, & Meyer, 2006). From there, seven crater lakes (Apoyeque, Apoyo, As. León, As. Managua, Masaya, Tiscapa and Xiloá) have been colonized (K. R. Elmer et al., 2014; Kathryn R. Elmer, Lehtonen, Fan, & Meyer, 2013; Kathryn R Elmer, Lehtonen, & Meyer, 2009). In two of the crater lakes, Apoyo and Xiloá, six and four endemic species have been described, respectively (Barlow & Munsey, 1976; Geiger, McCrary, & Stauffer Jr, 2010; Recknagel, Kusche, Elmer, & Meyer, 2013; Stauffer Jr, McCrary, & Black, 2008; Stauffer Jr & McKaye, 2002). In Crater Lake As. Manuagua, another endemic species, *A. tolteca* , has been formally described (Recknagel et al., 2013), while species of the other crater lakes await formal description (why we included them here as 'populations').

Crater lake populations and sympatric species therein clearly form separate clusters using both RAD-sequencing data (Kautt, Machado-Schiaffino, & Meyer, 2018) and whole-genome data (Kautt et al., 2020). While all crater lake populations and species differ morphologically (Kathryn R. Elmer, Kusche, Lehtonen, & Meyer, 2010; Kautt et al., 2018), species assignment can be difficult, especially when specimens are young, and particularly for the sympatric species from crater lakes Apoyo and Xiloá. Therefore, methods to quickly genotype fish using genetic markers would give additional confidence for species assignments and allow identification of species also for juvenile fish. This is important for certain research questions including for example cohort analyses and unbiased frequency estimations. Moreover, several of these species are protected or live in protected environments where illegal fishing occurs. Cheap genotyping assays with a fast turnaround time might contribute to conservation monitoring.

The objectives of this study were therefore to (1) design a workflow to screen for suitable GB-RFLP markers for species and population assignment, (2) test *in silico* if those markers would allow unambiguous assignment and (3) to perform GB-RFLP assays on independent samples (i.e., samples that have been not used for the design of the markers in (1)) to test if the markers are suitable to assign species and populations (i.e., lakes of origin).

2. Materials and Methods

2.1 Study system, samples and genomic data

The Midas cichlid species complex encompasses a total of — based on genetic clustering (Kautt et al., 2020) — at least 19 genetically distinguishable populations that include the 13 described species (Fig. 1). These include five crater lakes with only one species (Lakes Apoyeque, As. León, As, Managua, Masaya, and Tiscapa), one crater lake with six species (Lake Apoyo), one crater lake with four species (Lake Xiloá), and the great lakes Managua and Nicaragua that both harbor the same set of two species (*A. citrinellus* and *A. labiatus* ; Fig. 1). To find species and population-specific markers, we used genotype calls from a previous analysis of 453 re-sequenced Midas cichlid genomes, including all aforementioned species and lake populations (Kautt et al., 2020). Tissue samples were collected during field trips to Nicaragua (years 2003–2018).

2.2 Screen for species and population specific variants

To screen for species- and population-specific markers, we estimated population differentiation (F_{ST}) for individual SNP loci using *vcftools* (Danecek et al., 2011) always comparing the respective focal population (ingroup) to all other samples (outgroup) (the workflow is also described in Fig. 2). For Crater Lakes Apoyo and Xiloá, we compared the focal species (ingroup) against all other sympatric species (outgroup). For the two great lake species, we used instead genome-wide-association (GWA) data for lip shape, which is a characteristic trait that differs between *A. citrinellus* and*A. labiatus* and correlates with F_{ST} in both lakes (Kautt et al., 2020). For each species and population, we extracted the 20 variants with the highest differentiation (F_{ST}) / genome-wide association (GWA).

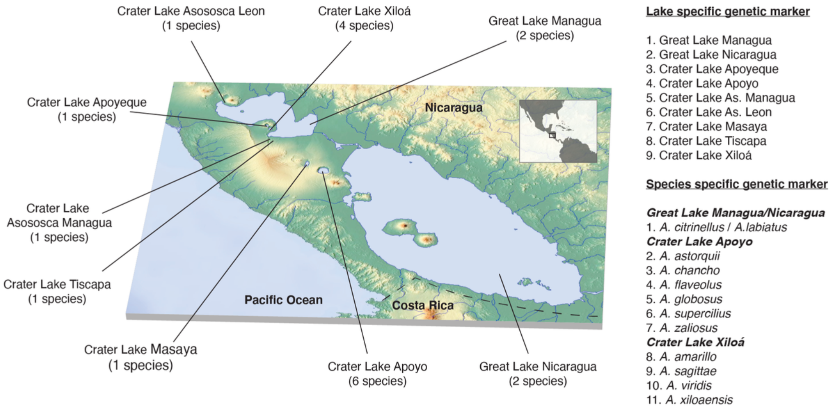

FIGURE 1 The Midas cichlid species complex includes fish inhabiting nine lakes (the two great lakes and seven crater lakes) and comprises 13 described species. In this study we designed genome-based PCR-RFLP (GB-RFLP) markers for population (lake specific genetic markers) and species assignment (species specific genetic markers).

2.3 Screen for RFLP alleles

To screen for RFLP alleles, we extracted 801 bp flanking the target variants (focal marker ± 400 bp) using samtools faidx from the *A. citrinellus* reference genome (Kautt et al., 2020). Genotypes for all loci were extracted from an in-house filtered statistically-phased variant call format (vcf) file (Kautt et al., 2020) using tabix. Nucleotide sequences and genotypes of variants (using *vcfR* (Knaus et al., 2020)) were loaded into R (R Development Core Team, 2019). For each locus, we generated a sequence with the alternative allele (alternative sequence) – for this initial analysis we did not use phased haplotypes. Next, we generated a list of the recognition sites of 237 commercially available restriction enzymes and performed *in silico* restriction digests of reference and alternative sequences for each locus. For each locus, we identified restriction enzymes cutting one of the alleles but not the other one and generated a list of all enzymes, including information on how many restriction sites could be found within the 801 bp reference sequence. For every locus, we then sorted the restriction sites by number of cutting sites within the 801 bp sequence.

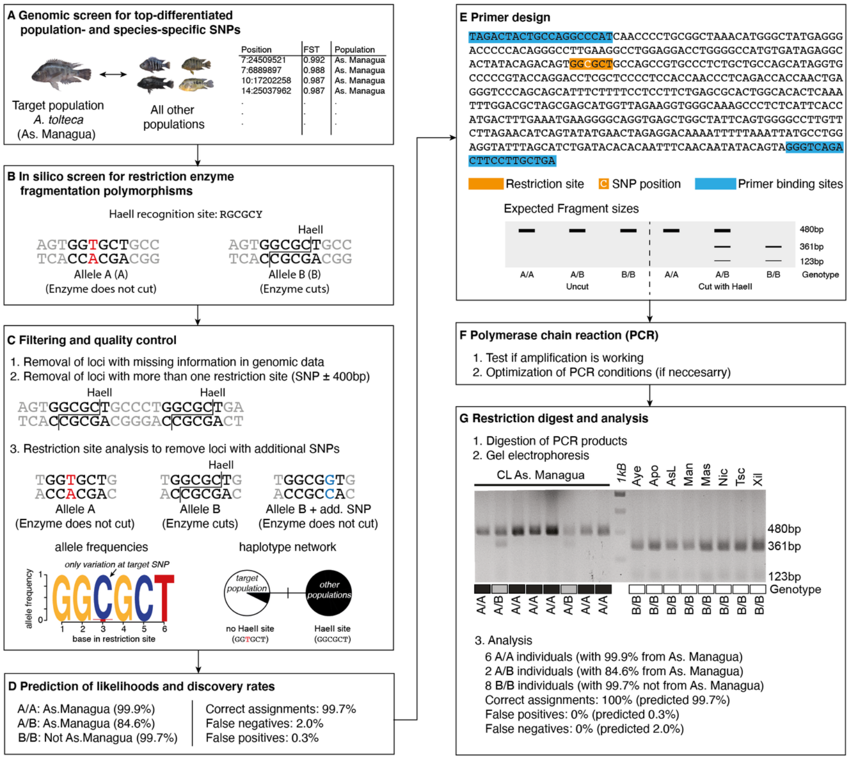
A Genomic screen for top-differentiated population- and species-specific SNPs
Target population
A. tolteca
(As. Managua)
All other populations
Position | FST | Population
7:24509521 | 0.992 | As. Managua
7:6889897 | 0.988 | As. Managua
10:17202258 | 0.987 | As. Managua
14:25037962 | 0.987 | As. Managua
B In silico screen for restriction enzyme fragmentation polymorphisms
HaeII recognition site: RGCGCY
HaeII
AGTGGTGCTGCC
TCACCACGACGG
Allele A (A)
(Enzyme does not cut)
AGTGGCGCTGCC
TCACCGCGACGG
Allele B (B)
(Enzyme cuts)
C Filtering and quality control
1. Removal of loci with missing information in genomic data
2. Removal of loci with more than one restriction site (SNP ± 400bp)
HaeII
HaeII
AGTGGCGCTGCCCTGGCGCTGA
TCACCGCGACGGGACCGCGACT
3. Restriction site analysis to remove loci with additional SNPs
TGGTGCTG
ACCACGAC
Allele A
(Enzyme does not cut)
HaeII
TGGCGCTG
ACCGCGAC
Allele B
(Enzyme cuts)
TGGCGGTG
ACCGCCAC
Allele B + add. SNP
(Enzyme does not cut)
allele frequencies
only variation at target SNP
allele frequency
1
0
1 2 3 4 5 6
base in restriction site
haplotype network
target population
other populations
no HaeII site (GGTGCT)
HaeII site (GGCGCT)
D Prediction of likelihoods and discovery rates
A/A: As.Managua (99.9%)
A/B: As.Managua (84.6%)
B/B: Not As.Managua (99.7%)
Correct assignments: 99.7%
False negatives: 2.0%
False positives: 0.3%
E Primer design
TAGACTACTGCCAGGCCCATCAACCCCTGCGGCTAAACATGGGCTATGAGGG
ACCCCCACAGGGCCTTGAAGGCCTGGAGGACCTGGGGCCATGTGATAGAGGC
ACTATACAGACAGTGGCGCTGCCAGCCGTGCCCTCTGCTGCCAGCATAGGTG
CCCCCGTACCAGGACCTCGCTCCCCTCCACCAACCCTCAGACCACCAACTGA
GGGTCCCAGCAGCATTTCTTTTCCTCCTTCTGAGCGCACTGGCACACTCAAA
TTTGGACGCTAGCGAGCATGGTTAGAAGGTGGGCAAAGCCCTCTCATTCACC
ATGACTTTGAAATGAAGGGGCAGGTGAGCTGGCTATTCAGTGGGGCCTTGTT
CTTAGAACATCAGTATATGAACTAGAGGACAAAATTTTTAAATTATGCCTGG
AGGTATTTAGCATCTGATACACACAATTTCAACAATATACAGTAGGGTCAGA
CTTCCTTGCTGA
Restriction site
C SNP position
Primer binding sites
Expected Fragment sizes
480bp
361bp
123bp
A/A
A/B
B/B
Uncut
A/A
A/B
B/B
Cut with HaeII
Genotype
F Polymerase chain reaction (PCR)
1. Test if amplification is working
2. Optimization of PCR conditions (if neccesarry)
G Restriction digest and analysis
1. Digestion of PCR products
2. Gel electrophoresis
CL As. Managua
1kB
Aye
Apo
AsL
Man
Mas
Nic
Tsc
Xil
480bp
361bp
123bp
Genotype
A/A
A/B
A/A
A/A
A/A
A/B
A/A
A/A
B/B
B/B
B/B
B/B
B/B
B/B
B/B
B/B
3. Analysis
6 A/A individuals (with 99.9% from As. Managua)
2 A/B individuals (with 84.6% from As. Managua)
8 B/B individuals (with 99.7% not from As. Managua)
Correct assignments: 100% (predicted 99.7%)
False positives: 0% (predicted 0.3%)
False negatives: 0% (predicted 2.0%)

FIGURE 2 Workflow of the whole genome re-sequencing-based design of PCR-RFLP markers (GB-RFLP). A. Markers were designed based on genetic differentiation (F_{ST}) or genome-wide genotype-phenotype association (GWA) data of a 453-genome dataset using pairwise ingroup–outgroup comparisons. B. Variants were screened for RFLPs with one allele (but not the other) being cut by a restriction enzyme. C. Target variants were filtered based on the presence of additional restriction sites and additional SNPs within the restriction site. Quality control was performed by plotting allele frequencies and haplotype networks. D. Likelihoods that a genotype corresponds to a population (or not) were calculated based on population-specific allele frequencies. Percentage of correct assignments together with false negative and false positive rates were based on bootstrapping of genotypes (informed by empirical allele frequencies in the genomic dataset). E. Primers were designed based on 801 bp sequences (core SNP +/- 400 bp) in a way that the restriction enzyme would generate two fragments with a ~1:2 length ratio. F. PCR conditions were tested and optimized. G. Restriction digest was performed on 8 ingroup samples and 5–8 outgroup samples. Genotypes were determined by visual inspection. These data were used to calculate the number of correct assignments as well as rates of false positives and negatives (as for the bootstrapping dataset in D).

2.4 Filtering and quality control

We filtered our genetic marker set in several ways. First, we removed variants that had missing data across the dataset and avoided markers that were on the same chromosome (to avoid that the markers are in linkage disequilibrium). Second, we selected 800 bp loci with a single restriction sites (or a maximum of two for a few). Third, we avoided genetic markers that had other variants in the flanking region (±5 bp). Based on the position of the restriction enzyme recognition sequence, we extracted the exact haplotypes for all sequenced genomes, estimated allele frequencies across the dataset and constructed haplotype networks (Fig. S1 and S2). Two RFLP loci had one additional SNP in the recognition sequence. However, these SNPs did not affect the RFLP. In one case the second SNP was fully linked to the target SNP (Fig. S2 L). In the other case there were two haplotypes that both were not cut by the enzyme and both more common in the outgroup (Fig. S1 O).

2.5 Primer design

Primers were designed using Primer3 (Sequences are summarized in Table S1). Briefly, we used the 801 bp sequence and designed the primers asymmetrically around the target SNP, i.e., that approximately one third (or two thirds) of the amplicon is 5' of the target SNP to avoid that the fragments have the same size after digestion. If the previous analysis indicated additional restriction sites in the 801 bp sequence we confirmed that the restriction sites are either outside of the amplicon, or at least do not complicate the detection of the RFLP (e.g., for the Lake Managua marker Chr3:33,260,056 the alleles resulted in 402, 106, and 50 bps or 448 and 106 bps long fragments).

DNA extraction and PCR

DNA was extracted as described previously (Kratochwil & Rijli, 2014). Briefly, we incubated fin or muscle tissue for 2h on a shaking incubator at 55°C using 200 µg/ml proteinase K (Sigma) in 500 µl extraction buffer (100 mM Tris–HCl pH 8.5, 200 mM NaCl, 5 mM EDTA, 0.2 % SDS) in 1.5ml tubes. Tubes were centrifuged at 12,000–16,000 × g. 500 µl isopropanol were added to the supernatant and centrifuged again. The supernatant was removed, and the pellet was washed with 500 µl of 70% ethanol. The pellet was air dried and diluted in 500 µl 10 mM Tris-Cl, pH 8.5. We did not use EDTA as it might affect restriction enzyme activity. PCRs were performed for eight samples (with the exception of a few cases where we did not have enough individuals) of the target population (ingroup) and 5–8 samples that did not belong to the target population (outgroup). For the outgroups, we tried to use representatives of each lake (for the lake comparisons, i.e., population comparisons) or other species living in sympatry (for the species comparisons). We did not include samples that were part of the previously-generated genome re-sequencing dataset. For PCRs, we used 2µl template and the standard protocol of DreamTaq DNA Polymerases (Thermo Fisher) with a total volume of 20µl, an annealing temperature of 58°C, 30 cycles and 30s extension time.

2.6 PCR-RFPL analysis

17 µl of the PCR amplicons were digested using the respective restriction enzymes (AccI, AciI, AleI, AlwNI, ApoI, AvaI, BbvI, BccI, BceAI, BclI, BsaAI, BsaBI, BsaI, BsiEI, BsmI, BspMI, BsrI, BstAPI, HaeII, HaeIII, HgaI, HinfI, HpaII, HphI, NcoI, NspI, PleI, PstI, PvuII, RsaI, Tth111I, XcmI; all from New England Biolabs) and recommended concentrations. PCR products were digested for 4h or overnight to avoid partial digestion and loaded on a 1% agarose gel. Genotypes were determined by visual inspection of gels and gel photographs (Fig. S3 and S4). Sometimes, despite long digestion time, we observed incomplete digestion. Genotypes were assigned as follows: 1) A lack of a smaller fragment was always interpreted as a homozygous genotype. 2) A band at the position of the uncut amplicon together with a second band was interpreted as a heterozygous genotype, but only if the uncut band was stronger in intensity (as the same number of longer-sized DNA molecules results in a stronger signal). 3) Complete lack of a band at the size of the uncut fragment or a band that had less intensity than the digested fragments were interpreted as a homozygous genotype for the alternative allele.

2.7 Estimation of RFLP marker quality

To estimate the reliability of individual markers, we first calculated discovery and false discovery rates based on allele frequencies. To do so, we first calculated allele frequencies by lake (lake comparisons) or sympatric species (species comparisons). Based on the allele frequency, we calculated expected genotype frequencies for all populations individually (assuming Hardy–Weinberg equilibrium) (Table S2). To evaluate the quality of the markers, we used these frequencies to calculate the chance that an individual with a particular genotype is from a particular population or not (without specifying which and assuming a 50:50 chance that the individual is from the focal population or not) (Table S2).

We used a bootstrap approach and randomly picked one million genotypes (i.e. individuals) from the ingroup (target population) and one million from the outgroups (again with an equal chance for each population/sympatric species to be picked) based on their relative frequencies and calculated how often a particular population/sympatric species would have been assigned correctly (correctly assigned), how often an ingroup individual would have been assigned to an outgroup (false negative) and how often an outgroup individual would have been assigned to the ingroup (false positive) (Fig. 3, Table S3). The same approach was then used based on our PCR-RFPL data (Table S3): False negatives were ingroup individuals that were incorrectly assigned as outgroup individuals, false positives were outgroup individuals that were incorrectly assigned as ingroup individuals. The proportion of correctly assigned individuals was calculated by taking the mean of the percentage of correctly assigned ingroup and the percentage of correctly assigned outgroup individuals (to make these estimates comparable to the estimates based on the bootstrapping dataset — some analyses were imbalanced with a different number of ingroup and outgroup individuals).

3. Results

3.1 GB-RFLP marker analysis for lake of origin assignment

As a first step, we chose RFLP markers that are specific for lakes (i.e., allowing to distinguish one population [ingroup] from all other populations [outgroup]) (Fig. 1 and Fig. 2). Although not all of them are (yet) described as separate species, genetic differentiation is very high among most of the lake populations — mainly because the crater lakes are completely isolated (and therefore not permitting gene flow) and because their founder populations were small (30–850 individuals), resulting in strong genetic drift (and thereby more alternatively fixed alleles) (Kautt et al., 2020). Most variants have likely been recruited from the standing genetic variation present in the source populations that was introduced into the crater lakes with the colonizers, as there was too little time for a substantial amount of*de novo* mutations to occur — all crater lakes have been colonized less than 5,000 years ago (Kautt et al., 2020; Kautt, Machado-Schiaffino, & Meyer, 2016; Kautt et al., 2018). The two great lakes Managua and Nicaragua are not isolated from each other, but intermittently connected by a river (Rio Tipitapa) resulting in limited ongoing gene flow (based on demographic models from (Kautt et al., 2020) for *A. citrinellus* approximately 1 in 12000 individuals per generation).

To assess the performance of markers, we calculated their predictive power (see detailed description in Materials and Methods section) based on bootstrapping using allele frequencies from a previous genomic study that was based on 453 re-sequenced genomes (Kautt et al., 2020) as well as GB-RFLP assays of additional samples in this study (that were not part of the resequenced genomes from the (Kautt et al., 2020) study, but morphologically assigned to lakes and species). In line with the discussed potential effect of the described demographic parameters (Kautt et al., 2020), RFLP markers were most powerful for crater lakes. For ten out of the twelve RFLP markers for Crater Lakes Apoyeque (Fig. S3E), Apoyo (Fig. S3G, H), As. Managua (Fig. S3I, J), As. León (Fig. S3K, L), Masaya (Fig. S3M, N), Tiscapa (Fig. S3O, P), and Xiloá (Fig. S3Q) more than 90% of the samples were correctly assigned to the lake (Table 1). The second Apoyeque marker (87%, Fig. S3F) and second Xiloá marker (69%, Fig. S3R) performed less well. For the great lakes Managua (both 62%) and Nicaragua (87% and 69%) the percentage of correctly assigned samples was — as expected — lower (Fig. S3A–D), as many variants are shared between the two great lakes or can be found at least in one of the crater lakes (as they have been colonized from the great lakes).

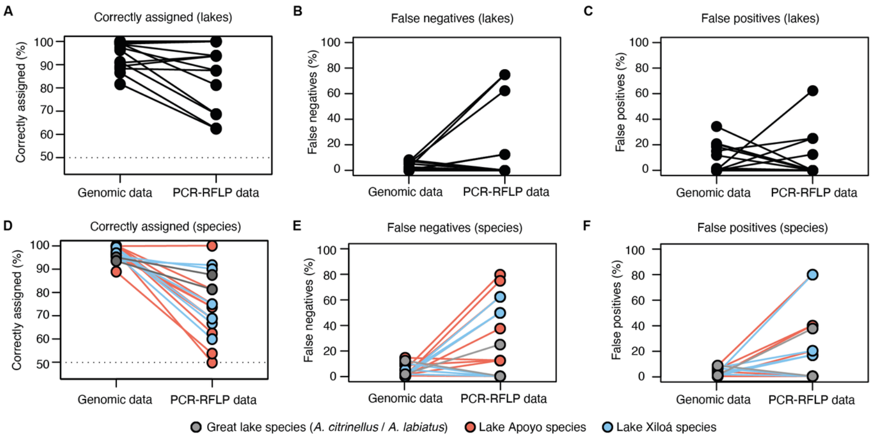

FIGURE 3 Comparison of the percentage of correctly assigned samples (A, D), false negatives (B, D) and false positives (C, E) using lake–specific GB-RFLPs (A–C) and species-specific GB-RFLPs. On the left side of the plots are always estimates based on bootstrapping data using allele frequencies from a previously generated genomic dataset are on the left, data from GB-RFLP analyses on the right. Color-code in D–E indicates the populations of Crater Lake Apoyo (red), Xiloá (blue) and the great lake species *A. citrinellus* and *A. labiatus*(grey).

Most markers showed similar accuracy as in the bootstrapping dataset that was based on allele frequencies in the respective lakes with 1/18 markers performing much worse (>20% fewer correctly assigned samples; Fig. S3R) in the RT-RFLP assay than in the bootstrapping dataset (Fig. 3A–C). We also tested whether a combination of two markers would improve accuracy. For two populations, Great Lake Nicaragua (69% instead of 62.5%) and Crater Lake Masaya (100% instead of 94%), we found a marginal improvement of correct assignments (Table S4).

TABLE 1 Tested RFLP markers, their location in the reference genome, used restriction enzyme, quality of the marker, and correctly assigned individuals (in %). Quality was assessed based on a combination of the predicted and tested number of correctly assigned specimens (++++: >99%, +++: >95%, ++: >90%, +: >80%, –: <80%). Ingroup means "within test population", outgroup "not within test population". For lake markers (above line) the outgroup contains all samples, for species markers (below line) only sympatric species within the same respective crater lake.

			Genotype (% correctly assigned)	**Genotype (% correctly assigned)**
Marker (Test population | coordinates)	**Enzyme**	**Marker quality**	**ingroup**	**outgroup**
Lake Managua | Chr3:33,260,056	HaeII		402/402 (99.8%) 550/402 (96.3%)	550/550 (73.8%)
Lake Managua | Chr21:22,155,441	BclI	-	400/400 (98.9%) 529/400 (88.2%)	529/529 (82.1%)
Lake Nicaragua | Chr11:21,205,346	BsrI	+	395/395 (98.9%) 544/395 (86.7%)	544/544 (85.5%)
Lake Nicaragua | Chr17:561,899	BsaAI	-	400/400 (99.2%) 551/400 (89.2%)	551/551 (88.9%)
Lake Apoyeque | Chr13:8,711,066	PleI	++	607/607 (100%) 607/367 (51.3%)	367/367 (100%)

			Genotype (% correctly assigned)	**Genotype (% correctly assigned)**
Lake Apoyeque | Chr11:1,877,014	PvuII	+	585/585 (99.7%) 585/363 (59.6%)	363/363 (99.9%)
Lake Apoyo | Chr18:5,677,584	HpaII	++++	523/523 (100%) 523/362 (59.4%)	362/362 (100%)
Lake Apoyo | Chr11:1,173,519	ApoI	++++	402/402 (100%) 533/402 (86.5%)	533/533 (100%)
Lake As. Managua | Chr7:6,889,897	HaeII	+++	480/480 (99.9%) 480/361 (84.6%)	361/361 (99.7%)
Lake As. Managua | Chr23:32,083,942	NcoI	+	196/196 (100%) 557/196 (100%)	557/557 (98.5%)
Lake As. León | Chr18:14,951,950	BsaI	++++	530/530 (100%)	405/405 (100%)
Lake As. León | Chr8:21,669,333	BceAI	++++	500/500 (100%)	500/188 (100%) 188/188 (100%)
Lake Masaya | Chr20:9,539,893	BccI	++	400/400 (100%) 553/400 (99.5%)	553/553 (84.7%)
Lake Masaya | Chr11:20,264,687	HaeIII	+	492/492 (100%) 492/363 (100%)	363/363 (82.4%)
Lake Tiscapa | Chr18:17,567,949	HphI	++++	411/411 (100%)	528/528 (100%)
Lake Tiscapa | Chr14:11,451,235	AvaI	++++	524/524 (100%) 524/401 (100%)	401/401 (99.8%)
Lake Xiloá | Chr3:13,299,069	BsaI	++++	543/543 (100%) 543/408 (97.1%)	408/408 (99.2%)
Lake Xiloá | Chr14:29,887,480	HgaI	-	506/506 (99.6%) 506/377 (63.4%)	377/377 (99.6%)
A. citrinellus / A. labiatus | Chr8:11,495,129	AccI	+	602/602 (98.5%)	602/351 (71%) 351/351 (99.8%)
A. citrinellus / A. labiatus | Chr8:11,381,985	NspI	+	348/348 (99.5%) 480/348 (57.4%)	480/480 (99.1%)
A. astorquii | Chr24:27,792,838	BsmI	-	370/370 (100%) 515/370 (96.8%)	515/515 (94.5%)
A. astorquii | Chr11:4,091,881	AlwNI	-	536/536 (99.4%) 536/389 (67%)	389/389 (99.2%)
A. chancho | Chr17:14,664,535	PstI	-	397/397 (99.9%)	520/397 (100%) 520/520 (100%)
A. chancho | Chr3:25,801,703	AciI	++++	112/112 (99.5%)	488/112 (100%) 488/488 (100%)
A. flaveolus | Chr15:22,314,952	ApoI	-	538/538 (95.2%)	538/185 (100%) 185/185 (100%)
A. flaveolus | Chr20:3,776,504	HinfI	-	524/524 (98.3%) 524/397 (74.7%)	397/397 (91.5%)
A. globosus | Chr3:19,696,124	BsaBI	-	536/536 (100%) 536/197 (100%)	197/197 (100%)

			Genotype (% correctly assigned)	**Genotype (% correctly assigned)**
A. globosus | Chr12:23,301,195	HgaI	-	382/382 (100%) 517/382 (100%)	517/517 (99.4%)
A. supercilius | Chr2:17,623,789	RsaI	-	583/583 (100%) 583/199 (96.4%)	199/199 (96.1%)
A. zaliosus | Chr24:14,867,912	BbvI	+	531/531 (99.9%)	531/413 (100%) 413/413 (100%)
A. amarillo | Chr6:811,192	BsiEI	-	514/514 (100%) 514/169 (86.8%)	169/169 (99.5%)
A. sagittae | Chr18:24,041,527	BstAPI	++	535/535 (99.3%) 535/171 (71.7%)	171/171 (97.6%)
A. viridis | Chr5:15,701,465	XcmI	+	175/175 (100%) 563/175 (96.2%)	563/563 (93.4%)
A. viridis | Chr16:24,619,817	BspMI	-	566/566 (99.8%) 566/185 (82.3%)	185/185 (97.8%)
A. xiloaensis | Chr4:5,998,299	Tth111I	-	512/512 (100%) 512/155 (100%)	155/155 (98.5%)
A. xiloaensis | Chr8:24,801,784	AleI	-	593/593 (99.8%)	593/391 (53.4%) 391/391 (99.9%)

3.2 GB-RFLP marker analysis for the assignment of sympatric crater lake species

Next, we designed markers to distinguish the sympatric species of crater lakes Apoyo (*A. astorquii* , *A. chancho* , *A. flaveolus* , *A. globosus* , *A. supercilius* and *A. zaliosus*) and Xiloá (*A. amarillo* , *A. sagittae* , *A. viridis* and *A. xiloaensis*). Although all sympatric species clearly form separate genetic clusters, the number of shared alleles is extensive due to ongoing gene flow and/or recent divergence (Kautt et al., 2016, 2018). Indeed, there are no alternatively fixed differences, but we show that there are some nearly fixed ones based on which we should reach 90 to 100% accuracy — however it is possible that these predictions might be limited because our genomic samples do not accurately enough reflect the actual population frequencies.

Overall, the quality of the species-specific RFLP markers (Fig. S4) was much lower than the lake specific markers (Fig. S3). The percentage of correctly assigned species ranged from 50% (equal to a random assignment) to 100%. Only two markers achieved values above 90% (*A. chancho* , Fig. S4F and *A. viridis* , Fig. S4O). In contrast to the lake-specific markers, species specific-markers showed a much lower accuracy in the GB-RFLP assay compared to the bootstrapping dataset with 12/16 samples having >20% less correctly assigned samples than in the bootstrapping dataset (Fig. 3D–F), which is substantially different from the lake-specific markers (1/18). The combination of markers did not improve the accuracy.

3.3 GB-RFLP marker analysis for the assignment of the great lake species A. citrinellus and A. labiatus

Lastly, we designed markers for the two great lake species. As genetic differentiation between these species is very low (0.015–0.02) compared to the crater lake or allopatric species pairs (0.07–0.5), we used SNPs that we have identified based on genome-wide association mapping for the species–defining trait: lip size (*A. labiatus* has thick, hypertrophic lips that *A. citrinellus* lacks) (Kautt et al., 2020; Machado-Schiaffino et al., 2017). As the phenotype is largely driven by a single locus, the variants were on the same chromosome (distance: 113kb). For the two species-specific markers that we designed, we predicted an accuracy of 92 and 99%, respectively. When tested, the GB-RFLP markers obtained similar accuracies of 87 and 81%, respectively (Fig. S4a, b and Fig. 3d–f). Interestingly, even though both loci are in proximity on the same chromosome, combination of both markers lead to an accuracy of 100%.

4. Discussion

Despite the increase in large-scale genomic data, PCR-RFLPs are still widely used as diagnostic markers for the detection and species assignment of parasites (Pegg et al., 2016), disease-causing pathogens (Kato et al., 2019), microbiota (Baffoni et al., 2013), toxic dinoflagellates (Lozano-Duque, Richlen, Smith, Anderson, & Erdner, 2018) as well as animals using different tissue samples (Larraín, González, Pérez, & Araneda, 2019), scat samples (Mukherjee, Cn, Home, & Ramakrishnan, 2010) or environmental DNA (eDNA) (Clusa, Ardura, Fernández, Roca, & García-Vázquez, 2017). Once markers are identified it is a fast, cheap, and reliable technique, but the design of PCR-RFLP markers is usually time consuming, especially if many species and populations are being compared and/or highly differentiated markers are difficult to find. Here, we introduce a streamlined workflow to identify PCR-RFLPs from whole genome re-sequencing data (GB-RFLPs). We note that the same approach could be applied to RAD-seq, exome sequencing, or other forms of targeted genomic data (Fig. 2).

Our study yielded promising results for diverged populations from different lakes without ongoing gene flow, as represented by all seven Nicaraguan crater lakes. While populations could be assigned with more than 90% accuracy to two crater lakes (Apoyeque and As. Managua), our markers even yielded 100% assignment accuracy for populations of the remaining five crater lakes (Table S4). Results were less clear for populations with ongoing gene flow and/or large population sizes, in particular the Great Lakes Nicaragua and Managua, for which population-specific markers performed poorly (between 62 and 86% assignment accuracy, Table S4). This was not unexpected as we know that many alleles are shared between the great lakes and chances are high that alleles found in one of the great lakes can at least be found in one of the crater lakes that was colonized from this older source population. Therefore, although we could assign individual samples using whole-genome (Kautt et al., 2020) or RAD-seq data (Kautt et al., 2018), single- or double marker approaches are not sufficient to unambiguously differentiate between Lake Managua or Lake Nicaragua Midas cichlid populations. A similar problem can be observed for the species–specific markers for the species of Crater Lakes Apoyo and Xiloá. Also here, species clearly form pronounced clusters using whole-genome (Kautt et al., 2020) or RAD-seq marker sets (Kautt et al., 2016). Yet, particularly in the sympatric scenario, where speciation occurred within the last 5,000 years (Kautt et al., 2020) and in at least one case gene flow persists (Kautt et al., 2020; Kautt et al., 2018), there might be a strong ascertainment bias when focusing on single SNPs — as it has been intensively discussed for SNP datasets from humans (Clark, Hubisz, Bustamante, Williamson, & Nielsen, 2005). In line with this caveat, indeed species-specific markers, with a few exceptions (*A. chancho* and *A. viridis*), performed less reliably (12/14 markers have <90% correct assignments; Table S4). Interestingly, the genetic markers for the great lake species that show extremely low genetic differentiation (FST˜0.02) perform quite well (87% and 81% correctly assigned), particularly when combined (100% correctly assigned) (Fig. S4). This can be explained by the different approach that was taken here. We designed markers based on the cognizant of our prior knowledge of the genomic basis of the species-defining trait of *A. labiatus* : hypertrophied, thick lips. As the trait and the underlying associated SNPs (lip size variation links to only a single locus in most populations; (Kautt et al., 2020)) are almost alternatively fixed between these species, the marker seem to be most powerful to reliably assign species. While signals for gene flow between*A. labiatus* and *A. citrinellus* can be detected in most of the genome, this is not true for the lip locus on chromosome 8, where also the genetic markers are located.

Based on our results, we conclude that the design of markers based on whole genome data is a powerful approach in an effort to distinguish clearly differentiated species or populations or rare cases where we have loci with high local differentiation that can be used as markers. For populations with ongoing gene flow or instances where the population constitutes the source population (both applies for Great Lakes Nicaragua and Managua) the single/double-GB-RFLP marker approach performs poorly — likely because our genomic samples that we used for the design of the markers gives only an estimate of the 'true' population allele frequencies (i.e., markers that seem perfect based on our limited genomic data are in reality not markers that can unambiguously differentiate populations). The same is true for sympatric species (Crater Lake Apoyo and Xiloá) without localized differentiation (as opposed to differentiation found between *A. labiatus* and*A. citrinellus*). To make reliable species identification possible, multi-marker assays might be necessary for some instances. These would likely not require the complete set of markers found via RAD-seq or WGS analyses but could be applied with a selected set of markers. Here, one approach would be to use those SNPs that load most heavily on the first principal components of Crater Lakes Apoyo and Xiloá (based on Kautt et al., 2020; Kautt et al., 2016, 2018) thereby giving most power to distinguish the sympatric species. Such very cost-effective targeted multi-SNP genotyping panels have been used, for example, for 217 SNPs to assign salmons to particular populations (Aykanat, Lindqvist, Pritchard, & Primmer, 2016) and might be an excellent approach, also for the Midas cichlid system. Lastly, this set of RFLPs is now available as a resource for conservation purposes to for example identify individual samples on fish markets, but also for cohort and mark-recapture studies. This study therefore also presents a workflow how use genomic resources for the generation of applicable low-budget approaches for species assignment. Our study therefore introduces a new methodological approach for such an effort, as implementation of approaches that can help 'real-world conservation issues' often fail as previously discussed (Shafer et al., 2015).

In summary, we tested a set of 36 PCR-RFLP loci that we designed based on whole genome re-sequencing data to genetically assign Midas cichlid species and populations. While our analyses reveal limitations for the assignment of species and populations with ongoing gene flow and/or extreme recent divergence, genome-based designed PCR-RFLPs (GB-RFLP) have great benefits when populations with robust genome-wide (allopatric populations) or local differentiation (*A. citrinellus* and*A. labiatus*) have to be identified.

Acknowledgements

The authors want to thank Sylvia Finkernagel for her crucial technical contributions. The authors also wish to thank the animal research facility (TFA) of the University of Konstanz. This work was supported by a European Research Council Advanced Grant (ERC "GenAdap" 293700) to AM and grants from the Deutsche Forschungsgemeinschaft to AM and CFK (KR 4670/2-1, KR 4670/4-1), as well as a grant from the Baden-Württemberg Foundation to CFK.

Conflict of interest

The authors declare that there are no conflict of interests.

Authors contributions

Sina Rometsch and Claudius Kratochwil collected data. Claudius Kratochwil and Andreas Kautt conducted the analysis. Claudius Kratochwil wrote the manuscript. All authors contributed to the manuscript. Axel Meyer and Claudius Kratochwil designed and coordinated the study.

Data Accessibility statement

All data to reproduce these results are part of the submitted manuscript. The genomic data this study is based on has been previously described in Kautt et al. 2020 and has been deposited at DDBJ/ENA/GenBank under accession JACBYM000000000 and PRJEB38173.

References

Auton, A., Abecasis, G. R., Altshuler, D. M., Durbin, R. M., Abecasis, G. R., Bentley, D. R., . . . National Eye Institute, N. I. H. (2015). A global reference for human genetic variation. *Nature, 526* (7571), 68-74. doi:10.1038/nature15393

Aykanat, T., Lindqvist, M., Pritchard, V. L., & Primmer, C. R. (2016). From population genomics to conservation and management: a workflow for targeted analysis of markers identified using genome-wide approaches in Atlantic salmon Salmo salar. *Journal of Fish Biology, 89* (6), 2658-2679. doi:10.1111/jfb.13149

Baffoni, L., Stenico, V., Strahsburger, E., Gaggìa, F., Di Gioia, D., Modesto, M., . . . Biavati, B. (2013). Identification of species belonging to the Bifidobacterium genus by PCR-RFLP analysis of a hsp60 gene fragment. *BMC Microbiology, 13* (1), 149. doi:10.1186/1471-2180-13-149

Barlow, G. W., & Munsey, J. W. (1976). The red devil-Midas-arrow cichlid species complex in Nicaragua.

Barluenga, M., Stölting, K. N., Salzburger, W., Muschick, M., & Meyer, A. (2006). Sympatric speciation in Nicaraguan crater lake cichlid fish.*Nature, 439* (7077), 719-723. doi:10.1038/nature04325

Campbell, M. R., Vu, N. V., LaGrange, A. P., Hardy, R. S., Ross, T. J., & Narum, S. R. (2019). Development and Application of Single-Nucleotide Polymorphism (SNP) Genetic Markers for Conservation Monitoring of Burbot Populations. *Transactions of the American Fisheries Society, 148* (3), 661-670. doi:10.1002/tafs.10157

Clark, A. G., Hubisz, M. J., Bustamante, C. D., Williamson, S. H., & Nielsen, R. (2005). Ascertainment bias in studies of human genome-wide polymorphism. *Genome research, 15* (11), 1496-1502. doi:10.1101/gr.4107905

Clusa, L., Ardura, A., Fernández, S., Roca, A. A., & García-Vázquez, E. (2017). An extremely sensitive nested PCR-RFLP mitochondrial marker for detection and identification of salmonids in eDNA from water samples.*PeerJ, 5* , e3045. doi:10.7717/peerj.3045

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., . . . Group, G. P. A. (2011). The variant call format and VCFtools. *Bioinformatics, 27* (15), 2156-2158. doi:10.1093/bioinformatics/btr330

Elmer, K. R., Fan, S., Kusche, H., Spreitzer, M. L., Kautt, A. F., Franchini, P., & Meyer, A. (2014). Parallel evolution of Nicaraguan crater lake cichlid fishes via non-parallel routes. *Nat Commun, 5* , 5168. doi:10.1038/ncomms6168

Elmer, K. R., Kusche, H., Lehtonen, T. K., & 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* (1547), 1763-1782. doi:doi:10.1098/rstb.2009.0271

Elmer, K. R., Lehtonen, T. K., Fan, S., & Meyer, A. (2013). Crater Lake Colonization by Neotropical Cichlid Fishes. *Evolution, 67* (1), 281-288. doi:https://doi.org/10.1111/j.1558-5646.2012.01755.x

Elmer, K. R., Lehtonen, T. K., & Meyer, A. (2009). Color assortative mating contributes to sympatric divergence of neotropical cichlid fish.*Evolution, 63* (10), 2750-2757.

Fennessy, J., Bidon, T., Reuss, F., Kumar, V., Elkan, P., Nilsson, M. A., . . . Janke, A. (2016). Multi-locus analyses reveal four giraffe species instead of one. *Current Biology, 26* (18), 2543-2549.

Geiger, M. F., McCrary, J. K., & Stauffer Jr, J. R. (2010). Description of two new species of the Midas cichlid complex (Teleostei: Cichlidae) from Lake Apoyo, Nicaragua. *Proceedings of the Biological Society of Washington, 123* (2), 159-173.

Kato, H., Gomez, E. A., Seki, C., Furumoto, H., Martini-Robles, L., Muzzio, J., . . . Hashiguchi, Y. (2019). PCR-RFLP analyses of Leishmania species causing cutaneous and mucocutaneous leishmaniasis revealed distribution of genetically complex strains with hybrid and mito-nuclear discordance in Ecuador. *PLOS Neglected Tropical Diseases, 13* (5), e0007403. doi:10.1371/journal.pntd.0007403

Kautt, A. F., Kratochwil, C. F., Nater, A., Machado-Schiaffino, G., Olave, M., Henning, F., . . . Meyer, A. (2020). Contrasting signatures of genomic divergence during sympatric speciation. *Nature, 588* , 106–111. doi:10.1038/s41586-020-2845-0

Kautt, A. F., 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 Genet, 12* (6), e1006157. doi:10.1371/journal.pgen.1006157

Kautt, A. F., Machado-Schiaffino, G., & Meyer, A. (2018). Lessons from a natural experiment: Allopatric morphological divergence and sympatric diversification in the Midas cichlid species complex are largely influenced by ecology in a deterministic way. *Evol Lett, 2* (4), 323-340. doi:10.1002/evl3.64

Knaus, B. J., Grunwald, N. J., Anderson, E. C., Winter, D. J., Kamvar, Z. N., & Tabima, J. F. (2020). R Package 'vcfR' v.1.10.0. *Cran R* .

Kratochwil, C. F., & Rijli, F. M. (2014). The Cre/Lox system to assess the development of the mouse brain. *Methods Mol Biol, 1082* , 295-313. doi:10.1007/978-1-62703-655-9_20

Larraín, M. A., González, P., Pérez, C., & Araneda, C. (2019). Comparison between single and multi-locus approaches for specimen identification in Mytilus mussels. *Scientific reports, 9* (1), 19714. doi:10.1038/s41598-019-55855-8

Lozano-Duque, Y., Richlen, M. L., Smith, T. B., Anderson, D. M., & Erdner, D. L. (2018). Development and validation of PCR-RFLP assay for identification of Gambierdiscus species in the Greater Caribbean Region.*Journal of Applied Phycology, 30* (6), 3529-3540. doi:10.1007/s10811-018-1491-5

Machado-Schiaffino, G., Kautt, A. F., Torres-Dowdall, J., Baumgarten, L., Henning, F., & Meyer, A. (2017). Incipient speciation driven by hypertrophied lips in Midas cichlid fishes? *Mol Ecol, 26* (8), 2348-2362. doi:10.1111/mec.14029

McKeown, N. J., Robin, J.-P., & Shaw, P. W. (2015). Species-specific PCR-RFLP for identification of early life history stages of squid and other applications to fisheries research. *Fisheries Research, 167* , 207-209.

Mukherjee, S., Cn, A., Home, C., & Ramakrishnan, U. (2010). An evaluation of the PCR-RFLP technique to aid molecular-based monitoring of felids and canids in India. *BMC Research Notes, 3* (1), 159. doi:10.1186/1756-0500-3-159

Nater, A., Mattle-Greminger, M. P., Nurcahyo, A., Nowak, M. G., De Manuel, M., Desai, T., . . . Roos, C. (2017). Morphometric, behavioral, and genomic evidence for a new orangutan species. *Current Biology, 27* (22), 3487-3498. e3410.

Ota, M., Fukushima, H., Kulski, J. K., & Inoko, H. (2007). Single nucleotide polymorphism detection by polymerase chain reaction-restriction fragment length polymorphism. *Nature protocols, 2* (11), 2857-2864. doi:10.1038/nprot.2007.407

Pegg, E., Doyle, K., Clark, E. L., Jatau, I. D., Tomley, F. M., & Blake, D. P. (2016). Application of a new PCR-RFLP panel suggests a restricted population structure for Eimeria tenella in UK and Irish chickens. *Veterinary Parasitology, 229* , 60-67. doi:https://doi.org/10.1016/j.vetpar.2016.09.018

Piertney, S. B. (2016). High-Throughput DNA Sequencing and the Next Generation of Molecular Markers in Wildlife Research. In R. Mateo, B. Arroyo, & J. T. Garcia (Eds.), *Current Trends in Wildlife Research* (pp. 201-223). Cham: Springer International Publishing.

R Development Core Team. (2019). *R: A language and environment for statistical computing. R Foundation for Statistical Computing* . Vienna, Austria.

Recknagel, H., Kusche, H., Elmer, K. R., & 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* (4).

Shafer, A. B. A., Wolf, J. B. W., Alves, P. C., Bergström, L., Bruford, M. W., Brännström, I., . . . Zieliński, P. (2015). Genomics and the challenging translation into conservation practice. *Trends in ecology & evolution, 30* (2), 78-87. doi:https://doi.org/10.1016/j.tree.2014.11.009

Stauffer Jr, J. R., McCrary, J. K., & Black, K. E. (2008). Three new species of cichlid fishes (Teleostei: Cichlidae) from Lake Apoyo, Nicaragua. *Proceedings of the Biological Society of Washington, 121* (1), 117-129.

Stauffer Jr, J. R., & McKaye, K. (2002). Descriptions of three new species of cichlid fishes (Teleostei: Cichlidae) from Lake Xiloá, Nicaragua. *Cuadernos de Investigación de la UCA, 12* , 1-18.

Torres-Dowdall, J., & Meyer, A. (in press). Sympatric and allopatric diversification in the adaptive radiation of Midas cichlids in Nicaraguan lakes. In M. E. Abate & D. L. G. Noakes (Eds.), *The Behavior, Ecology and Evolution of Cichlid Fishes: A Contemporary Modern Synthesis* (Vol. 40): Springer Netherlands.