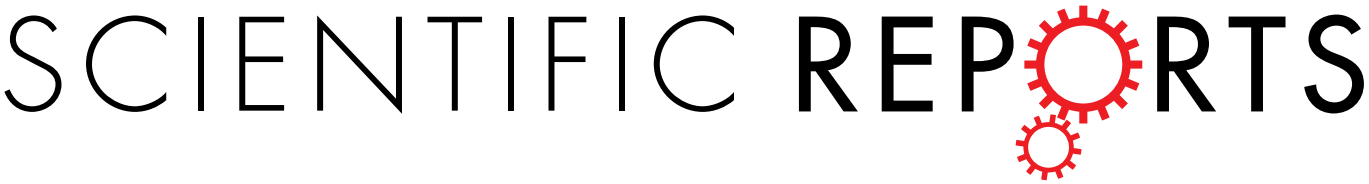

OPEN

Multilocus DNA barcoding – Species Identification with Multilocus Data

Junning Liu[1,2,3], Jiamei Jiang[1,2,3], Shuli Song[1,2,3], Luke Tornabene[4], Ryan Chabarria[5], Gavin J. P. Naylor[6] & Chenhong Li[1,2,3]

Received: 4 July 2017
Accepted: 20 November 2017
Published online: 30 November 2017

Species identification using DNA sequences, known as DNA barcoding has been widely used in many applied fields. Current barcoding methods are usually based on a single mitochondrial locus, such as cytochrome c oxidase subunit I (COI). This type of barcoding method does not always work when applied to species separated by short divergence times or that contain introgressed genes from closely related species. Herein we introduce a more effective multi-locus barcoding framework that is based on gene capture and "next-generation" sequencing. We selected 500 independent nuclear markers for ray-finned fishes and designed a three-step pipeline for multilocus DNA barcoding. We applied our method on two exemplar datasets each containing a pair of sister fish species: *Siniperca chuatsi* vs. *Sini. kneri* and *Sicydium altum* vs. *Sicy. adelum*, where the COI barcoding approach failed. Both of our empirical and simulated results demonstrated that under limited gene flow and enough separation time, we could correctly identify species using multilocus barcoding method. We anticipate that, as the cost of DNA sequencing continues to fall that our multilocus barcoding approach will eclipse existing single-locus DNA barcoding methods as a means to better understand the diversity of the living world.

DNA barcoding has been very successfully employed in many applied fields, ranging from routine species identification[1–3], to discovery of cryptic species[4,5], tracking of invasive species[6–8], conservation, and community ecology[9–12]. The mitochondrial cytochrome c oxidase subunit I gene (COI) has a good amount of variation and is easy to amplify using PCR based approaches in most animal groups[13–16]. It has become the most commonly used marker for animal DNA barcoding since it was first proposed more than a decade ago[13]. In most cases, single-locus (COI) DNA barcoding results in successful species identification. For example, a success rate close to 100% were reported for Germany herpetofauna[17], more than 90% for Chinese rodents[18], more than 80% for freshwater fishes of the Congo basin[19,20], and 100% for mosquitoes[21]. However, the success rate of species identification was low for species complexes with gene flow[17] or where species had only recently diverged[22].

In order to use barcoding for species identification, within species variation must be less than between species variation. This generates a "break" in the distribution of distances that is referred to as the "barcoding gap". Indeed one of the common causes of barcoding failure occurs when differences in demography eliminate the barcoding gap, because intra-specific differences are greater than inter-specific differences for the clades being compared. To an extreme, two individuals could have the same COI sequence, while being distinctly different species. Shared COI haplotypes have been reported in different species of spiders[23], birds[24] and fishes[25]. The single-locus barcoding is prone to misidentification when different species share haplotypes.

Although haplotypes at a single locus, such as COI can be shared between two species, it is unlikely that individuals of two species share alleles across multiple independent genes. Accordingly, multilocus data should perform better for species identification than any single locus could. Dowton *et al.*[26] proposed "next-generation" DNA barcoding based on multilocus data in which they incorporated multispecies coalescent species delimitation. They analyzed *Sarcophaga* flesh flies with two loci, mitochondrial COI and nuclear carbomoylphosphate synthase (CAD), and found out that their coalescent-based *BEAST/BPP approach was more successful than standard barcoding method[26]. However, Collins and Cruickshank[27] reanalyzed Dowton *et al.*'s data and showed that standard single locus (COI) barcoding method could achieve the same accuracy as the new multilocus

[1]Shanghai Universities Key Laboratory of Marine Animal Taxonomy and Evolution, Shanghai, 201306, China. [2]Key Laboratory of Exploration and Utilization of Aquatic Genetic Resources, Shanghai Ocean University, Ministry of Education, Shanghai, 201306, China. [3]National Demonstration Center for Experimental Fisheries Science Education (Shanghai Ocean University), Shanghai, 201306, China. [4]School of Aquatic and Fisheries Sciences, University of Washington, Seattle, WA, 98195, USA. [5]College of Science & Engineering, Texas A&M University – Corpus Christi, Corpus Christi, TX, 78412-5806, USA. [6]University of Florida, Gainesville, FL, 3261, USA. Correspondence and requests for materials should be addressed to C.L. (email: chli@shou.edu.cn)

framework did if an optimized distance threshold was applied[28–31]. The experiment of Dowton *et al.*[26] seems unsuccessful, but the likely reason for this is that the data they used was not challenging enough for standard single-locus barcoding methods, because there was only one unidentifiable individual that was more divergent from its closest putative conspecific than the optimized threshold[27]. The other reason is that only a single nuclear gene was used in their study, thus provided little additional information[27].

In the past it has been challenging to obtain sequences from sufficient independent nuclear loci from a broad taxonomic group to make multilocus DNA barcoding effective. Tools for finding thousands of nuclear gene markers[32–34] and collecting their sequences through cross-species gene capture and next-generation sequencing are now available[35], providing an opportunity to rigorously test the power of multilocus DNA barcoding. In this study, we screened for hundreds of nuclear gene markers for ray-finned fish and developed a three-step procedure for species identification. We tested our multilocus DNA barcodes in both empirical and simulated data. Our goal is to develop a multilocus barcoding approach for identifying species that are indistinguishable based on the current DNA barcoding method.

Results

We first investigated effect of increasing number of loci on species discrimination using empirical data (between *Siniperca chuatsi* and *Sini. kneri* and between *Sicydium altum* and *Sicy. adelum*). We subsequently estimated the population parameters, gene flow and divergence time for both pairs of species. Guided by the patterns seen in the empirical data, we simulated sequences with different splitting times and migration rates, and explored the effect of divergence time and gene flow on the success rate of species identification over a broader range of the relevant parameter space. Finally, we selected 500 nuclear markers for ray-finned fishes, designed a three-step pipeline for multilocus DNA barcoding and tested the new method on species identification.

Species discrimination using empirical data. We have developed 4,434 single-copy nucleotide loci for ray-finned fishes, and tested them in 83 species (33 families and 11 orders), covering major clades of ray-finned fishes[36]. Those markers have few missing data in the taxa tested, showing promise for their deployment in phylogenetics and population genetic analyses. We adopted those 4,434 loci as candidate barcoding markers in order to further optimize a subset of universal markers for all ray-finned fishes. We choose loci that could be readily captured and sequenced across taxa, and that were variable based on their average p-distance values among taxa.

Some of the most challenging instances for DNA barcoding occur when taxa are recently diverged or when gene flow exists between closely related species, or both. In an effort to design a rigorous barcoding scheme, we picked empirical study systems that would involve both challenges. The first involved sinipercid fishes, a family of fishes containing two genera, 9 to 12 species depending on the authority referenced[37–40]. Among them, two sister species, *Siniperca chuatsi* and *Sini. kneri* have distinct morphological characters, such as number of plyoriccaecum, ratio between eye length and head length[39], but they are not distinguishable using mitochondrial control region sequences[41]. These two sister species are allopatric in most of their distribution regions[39,40], so the reason for unsuccessful species identification in these sister species is likely due to their recency of speciation[41]. The other group of fishes that we checked is *Sicydium*. *Sicydium* is a group of diadromous gobies native to fast-flowing streams and rivers of the Americas (Central America, Mexico, Cocos Island, the Caribbean, Colombia, Ecuador and Venezuela) and Africa. There are two syntopic species, *Sicy. altum* and *Sicy. adelum* that could be separated according to distinct dental papillae and other morphological characters[42], but they are indistinguishable using mitochondrial or nuclear genes[43]. Because these two closely related species are frequently found together[42], it is possible that they have been subject to interspecific gene flow which would account for the high degree of genetic similarity between them. These two pairs of sister-species were used as test cases to evaluate how gene flow and shallow divergence times might affect species discrimination and identification based on multilocus barcoding.

After all loci with missing taxa were excluded, 2,586 loci were retained for *Siniperca*. The intra- and interspecific p-distances between five individuals of *Sini. chuatsi* and five *Sini. kneri* using different numbers of nuclear loci or COI are shown in Fig. 1. The intraspecific p-distance (red) calculated using one locus or a small number of loci overlap with interspecific p-distance (blue). There is no barcoding gap separating the intra- and interspecific distances. Intraspecific distances did not become distinguishable from interspecific distances until more than 90 loci were used. The gap separating the intra- and interspecific distance increased as more loci were added, but had little effect after 400 loci were used. The variance of the intra- and interspecific p-distance decreased when more loci were included in calculating the p-distance. The p-distance calculated on COI sequences had overlapped intra- and interspecific values, so that the mean intraspecific distance was 0.0462 (0.0025 – 0.1542) and the mean interspecific distance was 0.0539 (0.0025–0.1347) (Fig. 1).

Similar p-distance calculations on *Sicy. altum* and *Sicy. adelum* resulted in a different pattern than the one observed for *Siniperca*. The intra- (red) and interspecific (blue) p-distances in *Sicydium* were always mixed together, no matter how many loci were included in the analysis. The variance of intra- and interspecific p-distance decreased when more loci were included. The intra- and interspecific p-distances calculated using COI also were indistinguishable (Fig. 2).

The success rate of identification was low (0.412) in *Siniperca* when only one locus was used based on "all species barcodes" criterion[44] with two hundred trials, but it rose up quickly and reached 1.0 after more than 90 loci were added to the dataset (green dots, Fig. 3; Supplementary Table S1). The identification success rate was zero in *Sicydium* according to the "all species barcodes" criterion, no matter how many loci were included in the analysis (red triangles, Fig. 3). We also applied the COI barcoding approach with an optimized threshold[28]. The success rate of species identification using COI was zero in both *Siniperca* and *Sicydium*.

Population parameters inferred for the two species pairs. To investigate the difference seen in the results of the *Siniperca* and *Sicydium* analyses, we explored some of the population attributes associated with

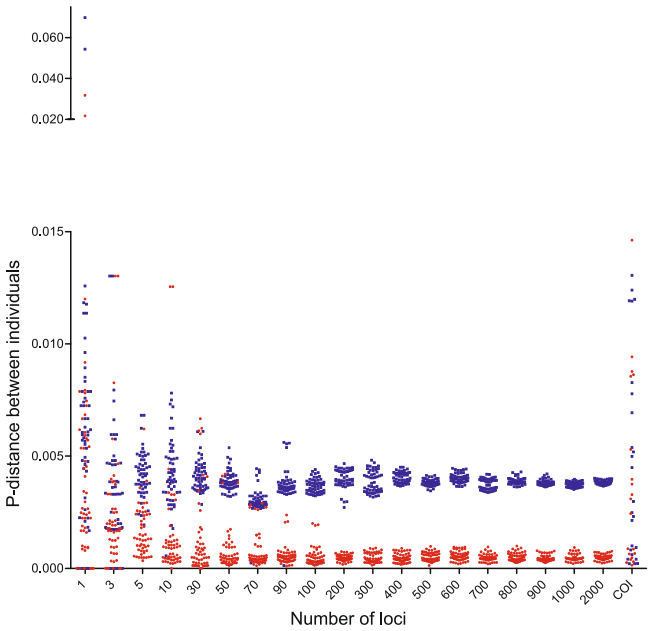

Figure 1. Intra- (red) and interspecific (blue) p-distance of *Siniperca chuatsi* and *Sini. kneri* calculated using different number of nuclear loci or COI gene. Scale of distances larger than 0.020 was reduced to fit all data points in the art board.

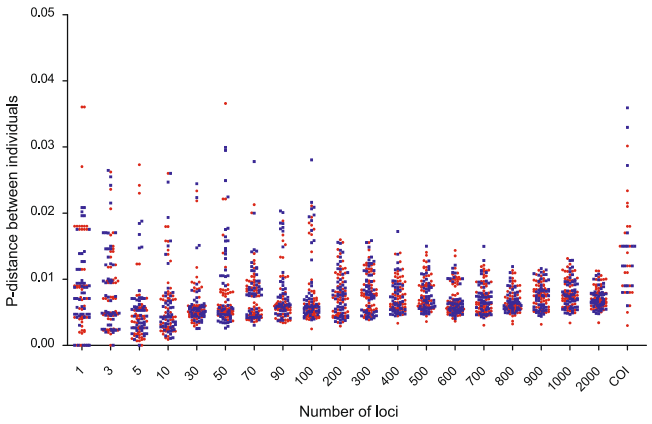

Figure 2. Intra- (red) and interspecific (blue) p-distance of *Sicydium altum* and *Sicy. adelum* calculated using different numbers of nuclear loci or COI gene.

each of these two groups. Structure analysis showed that K equaled to 2 had the highest probability when analyzing the two *Siniperca* species (Supplementary Fig. S1), but the two *Sicydium* species were indistinguishable (Supplementary Fig. S1). The divergence time between *Sini. chuatsi* and *Sini. kneri* was estimated as $t_0 = 1.754$, which would be equal to ~800,000 generations if we assume an average locus size of 300 bp, a generation time of 2–3 years for *Siniperca* and a substitution rate of 2.22×10^{-9} per site per year[45]. Gene flow from *Sini. chuatsi* to *Sini. kneri* was 0.157 (not significant by LLRtest), but gene flow from *Sini. kneri* to *Sini. chuatsi* was highly significant, 0.640 ($p < 0.001$). The divergence time between *Sicy. altum* and *Sicy. adelum* was estimated as $t_0 = 0.003195$, which was not significantly different from zero ($HPD95_{Lo} = 0$). Gene flow from *Sicy. altum* to *Sicy. adelum* was 0.494, and gene flow from *Sicy. adelum* to *Sicy. altum* was 0.502.

Simulation results. To explore the effect of divergence time and gene flow on the success rate of species identification, we conducted a series of simulations using twenty thousand loci for two species with a range of splitting times and migration profiles. Five sequences from each species were sampled to calculate species identification success rate. Different number of simulated loci were randomly picked and used to identify species. The identification success rate rose with increasing number of loci included in the analyses in all scenarios (Fig. 4). When there was no migration between the two simulated species, the identification success rate increased with splitting time (Fig. 4a). The simulation with a splitting time of 1,000 generations had the worst identification success rate, only 0.111 even with 1,000 loci used (green circle, Fig. 4a; Supplementary Table S2). The samples

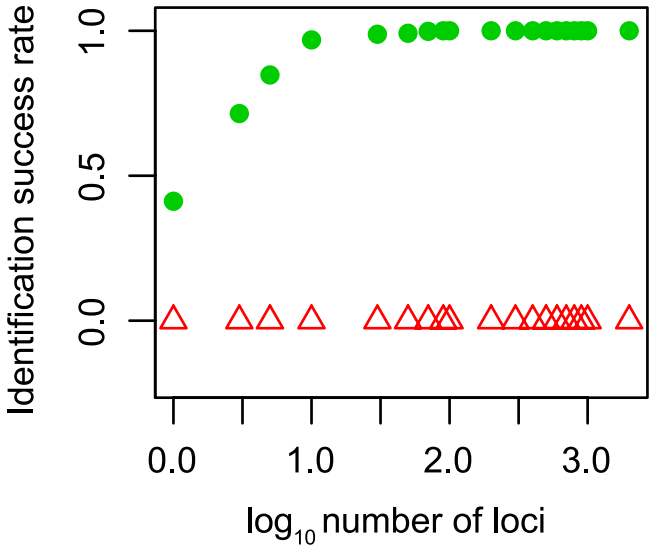

Figure 3. The relationship between number of loci used and success rate of identification between *Siniperca chuatsi* and *Sini. kneri* (green dots), and between *Sicydium altum* and *Sicy. adelum* (red triangles).

with a splitting time of 10,000 generations had low success rates with a small number of loci used, but rose to 1 when more than 400 loci were added to the analyses (blue triangles, Fig. 4a). The samples with a splitting time of 100,000 generations had a success rate of 1 when more than 10 loci were used (black crosses, Fig. 4a). Samples with a splitting time of 700,000 had success rate of 1 for all analyses (red line, Fig. 4a).

When gene flow was considered, high gene flow worked in concert with shallow divergence time to reduce the identification success rate (Fig. 4b-d). A migration rate of 0.0001 (per gene per generation) always led to the worst success rate, and failed to reach a success rate of 1.0 even when all 1,000 loci were used in analyses (green circles, Fig. 4b–d; Supplementary Tables S3–S5). At a migration rate of 0.00001 (blue triangles, Fig. 4b–d) or a migration rate of 0.000001 (black crosses, Fig. 4b–d), the identification success rate improved quickly with increasing number of loci (Fig. 4b–d; Supplementary Tables S3–S5). When the divergence time was greater than 100,000 generations and gene flow was lower than 0.00001, the identification success rate reached 1.0 when more than 90 loci were added to the analysis (Fig. 4c–d; Supplementary Tables S4–S5).

To test whether the length of sequence or the number of loci was the key for success in species identification, we simulated a single locus with increasing size matching the total length of multiple loci. We found that increasing the length of a single locus from 300 bp to 9,000 bp improved the success rate slightly, but the success rate did not change when longer sequences were used (Fig. 5 red circles; Supplementary Table S6). In contrast, concatenating more independent loci with the same total length as the single locus continuously improved the identification success rate, until it reached one (Fig. 5 blue triangles; Supplementary Table S6).

Multilocus DNA barcoding using empirical data. Based on the results from p-distance and species identification analyses of simulated and empirical data, we decided to pick 500 loci for multilocus DNA barcoding. First, we filtered the 4,434 markers developed for all ray-finned fishes and kept 750 loci with the lowest number of missing taxa. Next, we sorted the 750 loci by their average p-distance and picked from them 500 independent loci with large p-distances. This design was implemented both to minimize missing data when applying to ray-finned fishes and to ensure that loci would be variable enough for multilocus DNA barcoding. Information describing the 500 loci is listed in Supplementary Table S7.

Three individuals, 839_3 (*Sini. kneri*), 839_6 (*Sini. kneri*), and 938_1 (*Sini. chuatsi*) were randomly selected. Each of the randomly picked individuals was used to simulate "a putatively unknown" query for identification. Firstly, the p-distance between the unknown query and the other sinipercids in the database was calculated (Supplementary Table S8). Secondly, based on the sorted list of p-distances, we selected five closely related taxa, including the query. For example, for 839_3, we used sequence data of 839_3, *Sini. kneri*, *Sini. chuatsi*, *Sini. undulata* and *Sini. obscura* to reconstruct a species tree, in which 839_3 was found to be sister to *Sini. kneri* (Supplementary Fig. S2). We then ran a BFD* test to delimitate the unknown query (839_3) with *Sini. kneri* using *Sini. chuatsi* as outgroup. The BFD* analyses correctly grouped 839_3 (*Sini. kneri*) with *Sini. kneri* (Table 1). The two other randomly picked samples, 839_6 (*Sini. kneri*), and 938_1 (*Sini. chuatsi*) were also correctly identified (Supplementary Tables S8 and S9).

DNA barcoding using only COI data was unsuccessful. In many cases, the closest taxa of the unknow samples were not their conspecifics either in the tree or measured by p-distances (Supplementary Fig. S3 and Table S10).

Effect of missing data on multilocus DNA barcoding. When all conspecifics were excluded from the database, the unknown query, 893_3 (*Sini. kneri*) was found to be closely related to its sister species *Sini. chuatsi* (Supplementary Fig. S4). The p-distances also indicated that the unknown was related to *Sini. chuatsi* (Supplementary Table S11). A species delimitation analysis was run with the BFD* method to test whether the

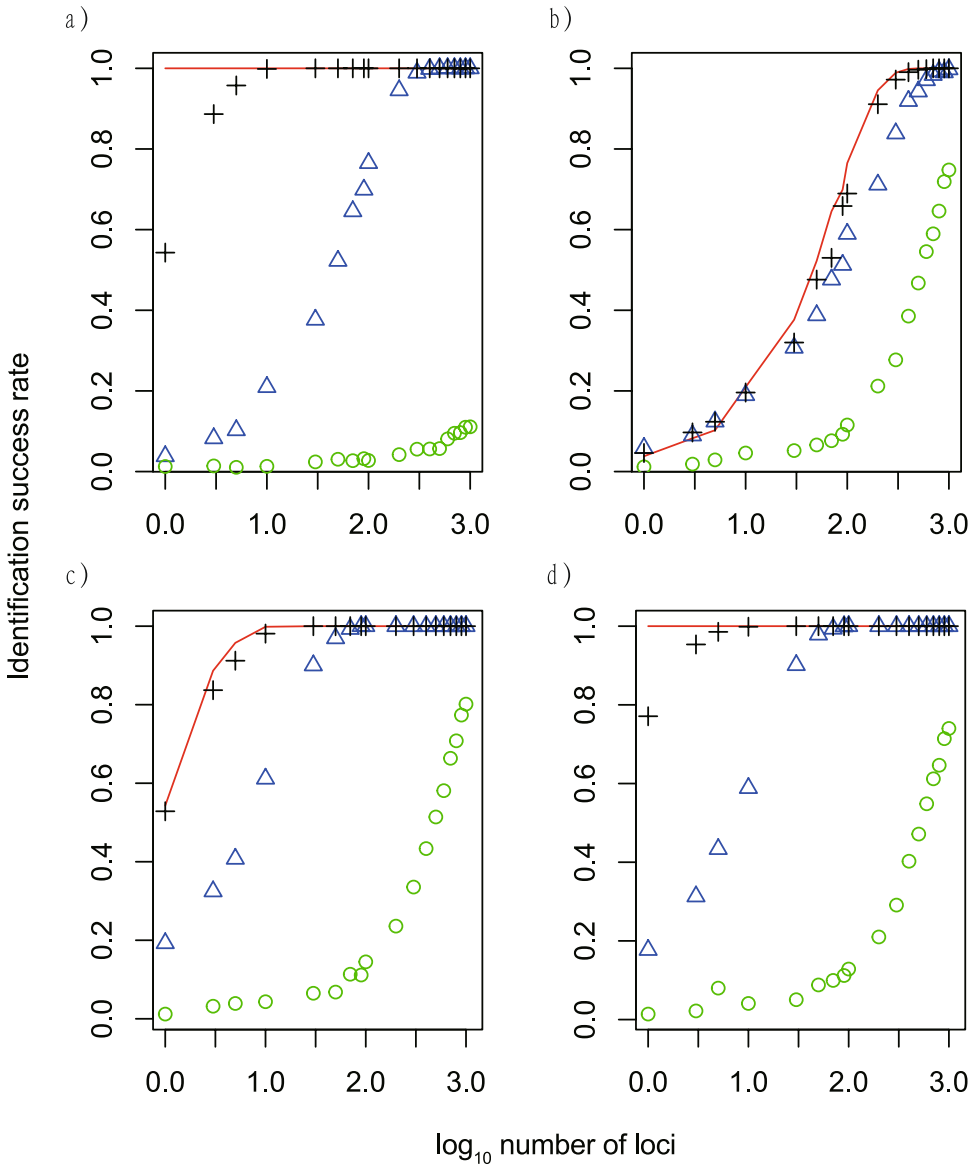

Figure 4. Identification success rate using simulated sequences under different scenarios. (**a**) migration rate equals zero and divergence time equals 700,000 generations (red line), 100,000 generations (black crosses), 10,000 generations (blue triangles), and 1000 generations (green circles); (**b**) divergence time equals 10,000 and migration rate equals 0 (red line), 0.000001 (black crosses), 0.00001 (blue triangles) and 0.0001 (green circles); (**c**) divergence time equals 100,000 and migration rate equals 0 (red line), 0.000001 (black crosses), 0.00001 (blue triangles) and 0.0001 (green circles); (**d**) divergence time equals 700,000 and migration rate equals 0 (red line), 0.000001 (black crosses), 0.00001 (blue triangles) and 0.0001 (green circles).

unknown should be assigned to *Sini. chuatsi* or not. The result of BFD* strongly support the unknown query as a separate species (2lnBF = 2,255.7; Table 1). In other tests, we keep the database intact, but excluded 20%, 30% and 50% of the loci from the unknown query (893_3 *Sini. kneri*). We still identified the unknown correctly using the multilocus DNA barcoding approach (Table 1).

Discussion

Our results demonstrated that the difference between species become more distinct when more independent loci are used. The intra- (red) and interspecific (blue) p-distance between individuals of *Sini. chuatsi* and *Sini. kneri* were largely overlapping when only COI gene or a few randomly picked nuclear gene were used to calculate the p-distance (Fig. 1). When more loci were added to the analyses, the intra- and interspecific distance became better separated. At 90 loci, a "barcoding gap" between the intra- and interspecific distance emerged. The variance of the intra- and interspecific distances also decreased as the number of loci used in the analyses increased. Based on these findings we conclude that the lack of an apparent barcoding gap between *Sini. chuatsi* and *Sini. kneri* using COI or a few nuclear genes is due to sampling error. Using more independent loci would likely improve the estimates of population parameters[46]. Similarly, more indepenent loci should improve precision of both the

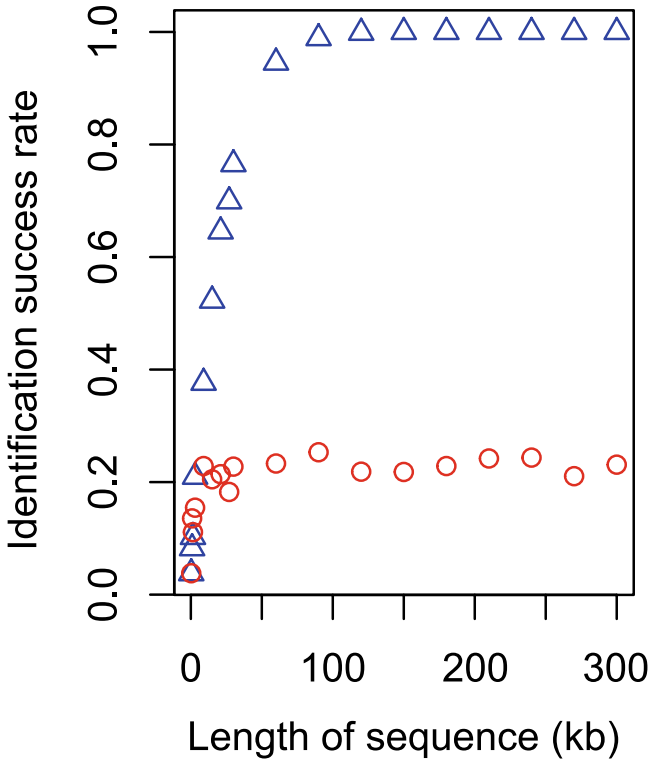

Figure 5. Comparison between success rates of species identification based on a single locus (red circles) and multiple loci (blue triangles). The length of the single locus equals the total length of multiple loci (300 bp each).

Data treatment	Model	Marginal likelihood	2lnBF
Using all data	Lumping 839_3 and *Sini. kneri*	−1575.80	20.62
	Splitting 839_3 and *Sini. kneri*	−1586.11	
Excluding conspecifics of 839_3	Lumping 839_3 and *Sini. chuatsi*	−2350.77	
	Splitting 839_3 and *Sini. chuatsi*	−1222.90	2255.7
Excluding 20% loci of the 839_3	Lumping 839_3 and *Sini. kneri*	−1467.41	26.54
	Splitting 839_3 and *Sini. kneri*	−1480.68	
Excluding 30% loci of the 839_3	Lumping 839_3 and *Sini. kneri*	−1247.44	22.12
	Splitting 839_3 and *Sini. kneri*	−1258.50	
Excluding 50% loci of the 839_3	Lumping 839_3 and *Sini. kneri*	−914.60	22.40
	Splitting 839_3 and *Sini. kneri*	−925.80	

Table 1. Results for species delimitation on unknown sample 893_3 (*Sini. kneri*) using BFD* based on all 500 nuclear loci, missing 20%, 30% and 50% of the 500 loci or missing conspecific of *Sini. kneri* in the database.

estimated intra- and interspecific genetic distance, resulting in increased discriminatory power (Fig. 3). The same patterns were obeserved in all of our simulated analyses, namely that the species identification success rate rose with increasing number of loci (Fig. 4). Interestingly, using longer genes instead of more genes did not improve species identification (Fig. 5).

Gene flow between sister species can cause problems that are similar to those caused by a lack of divergence. *Sini. chuatsi* and *Sini. kneri* were estimated split at around 800 thousand generations ago, with uni-directional introgression flowing from *Sini. kneri* to *Sini. chuatsi*, $m_{1>0} = 0.640$. Therefore, the lack of reciprocal monophyly or barcoding gap between *Sini. chuatsi* and *Sini. kneri* using COI or a few nuclear loci could, in fact, be caused by gene flow between the two species rather than the short divergence time originally hypothesized by us.

Sicydium altum and *Sicy. adelum* were estimated to have split very recently, $t_0 = 0.003195$. Bi-directional gene flow was estimated as 0.494 from *Sicy. altum* to *Sicy. adelum*, and 0.502 from *Sicy. adelum* to *Sicy. altum*. All of our analyses could not differentiate between *Sicy. altum* and *Sicy. adelum* genetically. Structure analysis (Supplementary Fig. S1), species identification and p-distance assessments (Figs 2 and 3) all indicated that *Sicy. altum* and *Socy. adelum* are indistinguishable. Accordingly, we suggest that the taxonomic status of *Sicy. altum* and *Sicy. adelum* be revisited by a more detailed morphological analysis.

It is difficult to tell whether gene flow or short species divergence time played a more prominent role in obstructing DNA barcoding. It has been reported that a considerable proportion of animal species do not form

monophyletic groups[47,48], but the causes for such patterns have not yet been fully explored. From the results of our empirical and simulated analyses, we conclude that when the splitting time between sister species was more than 100,000 generations old and the migration rate was lower than 0.00001, using multilocus DNA barcoding (with more than 90 loci) we could correctly determine the species status of unknown samples, whereas single-locus DNA barcoding suffered from lacking of power in species discrimination.

A suite of universal gene markers that could be used on a whole group of organisms is a prerequisite for multilocus DNA barcoding. Because of improvements in sequencing technology and the increasing number of publicly accessable genome data bases, more and more genome-scale markers have been developed for different group of organisms, such as turtles[49], birds[50], tapeworms[51], flower flies[52], plants[53], echinoderms[54], insects[55] and vertebrates[35]. Some of these markers can be applied across broad groups of organisms, whereas other have only been tested for restricted groups. We predict that obtaining suitable sets of markers for multilocus DNA barcoding will not be a limitation, but a lot of testing will need to be carried out across a broad range of taxa before an agreed set of common markers can be established for each major group of organisms.

Our pick of 500 markers for ray-finned fishes has been tested in major lineages of fishes (33 families and 11 orders). We chose markers that were found to be present in most groups of fishes and that were variable across groups. We recommend using them as standard multilocus DNA barcode markers for all ray-finned fishes. Our results indicate that more than 90 loci should be enough for species identification, but we advocate using the complete set of 500 loci, as there is almost no extra cost in capturing 500 rather than 90 loci. Additionally, targeting more loci provides insurance against missing data. We found that missing 20%, 30% and up to 50% loci in the unknown sample had no effect in identification success.

Other alternatives to collecting large datasets for DNA barcoding include genome skimming[56] and whole-chloroplast genome sequencing[57]. Genome skimming employs low-coverage shotgun sequencing of genomic DNA, which circumvents the need for PCR, avoiding the needs for univerals primers. Because genome skimming is unselective, it involves collecting a lot of data that ultimately is not used, but requires data storage and analysis resources. Low-coverage shotgun sequencing also yields a high proportion of missing data. Sequencing genomes of chloroplasts or other organelles is focused on a single long sequence, which tends to yield low success rate of species identification, as shown in our simulation.

Dowton *et al.*[26] proposed an pipeline integrating species tree reconstruction and species delimitation. They used Beast* to build a species tree[58], and took the species tree as the guide tree for delimitating species using BPP[59,60]. Our method is similar to the method of Dowton *et al.*[26]. We first screened the reference database for individuals from closely related speices based on p-distance between the unknow query and sequences in the database. We only choose four closely related species as potential conspecies or sister species. We think the number of species selected is enough for the current study, because our p-distance calculation was based on many independent loci, which reduced random error. The small number of selected species could also help to relieve computational burden associated with reconstructing the species tree in the second step. Using a combination of RAxML and ASTRAL program, we could reconstruct a specis tree of five taxa, four selected species plus the query in minutes using 500 loci. In the last step, we included only three taxa, one conspecific or sister species, one outgroup species and the unknown query for species delimiation using BFD*, which also saved computation time. Our multilocus barcoding approach is conceptually similar to the method of Dowton *et al.*[26]. However, we used many more loci (hundreds vs two), and focused on rooted trees with only three taxa, so it should be more powful and tractable than the method of Dowton *et al.* We anticipate that the computational burden associated with multilocus DNA barcoding will be further reduced as new algorithms are developed, to make multilocus barcoding a real-time tool.

Finally, from a practical standpoint, multilocus barcoding through target gene enrichment is efficient. We estimate around $90 for the total cost of capturing and sequencing 500 loci per sample, which is less than the cost of amplifying and sequencing 10 loci using the tranditional methods of PCR and Sanger sequening. The cost of target gene capture comprises: library prep, $50; RNA baits, $32; and sequencing, $8 per sample. The major costs are associated with the purchase of commercial RNA bait kits and the library preparation step, which can be lowered by purchasing kits in bulk and by using robots to automate library preparation. Finally, we selected 20 loci that have few missing data and long sequence length and recommend these for who want to use regular PCR and Sanger sequencing to collect multilocus data for species identification (Supplementary Table S12). These markers also can be used for phylogenetic study in the ray-finned fishes.

Materials and Methods

Taxa sampling, target gene enrichment, sequencing and reads assembly. We used the sequence data of the 4,434 loci of the sinipercids from Song *et al.*[61]. The samples included five *Coreoperca whiteheadi*, one *Sini. scherzeri*, five *Sini. obscura*, two *Sini. undulata*, three *Sini. roulei*, five *Sini. chuatsi* and five *Sini. kneri.*

For the goby study, nine *Sicy. altum* and seven *Sicy. adelum* were collected from Costa Rica. Total genomic DNA was extracted from fin clips using a Tissue DNA kit (Omega Bio-tek, Norcross, GA, USA) and the concentration of DNA was quantified using NanoDrop 3300 Fluorospectrometer (Thermo Fisher Scientific, Wilmington, DE, USA). The goby samples were enriched and sequenced for the same 4,434 loci. The amount of DNA used for library preparation was 1 μg for each sample. The DNA sample was first sheared to 250 bp using a Covaris M220 Focused-ultrasonicator™ (Covaris, Inc. Massachusetts, USA). A MYbaits kit containing baits for the 4,434 loci was synthesized at MYcroarray (Ann Arbor, Michigan, USA). The baits were designed on sequences of *Oreochromis niloticus* with 3 × tiling. Blunt-end repair, adapter ligation, fill-in, pre-hybridization PCR and target gene enrichment steps followed the protocol of cross-species gene capture[35]. The enriched libraries were amplified with indexed primers, pooled equimolarly and sequenced on a lane of Illumina HiSeq. 2500 platform with other samples. The raw reads were parsed to separate file for each species according to the indices on the adapter. Reads assembling followed the pipeline of Yuan *et al.*[51]. Mitochondrial COI gene of both the

sinipercids and the gobies was also amplified and sequenced using Sanger sequencing to compare COI barcoding with multilocus DNA barcoding using two pairs of primers (siniF: AACCAGCGAGCATCCATCTA and siniR: CAGTGGACGAAAGCAGCAAC for the sinipercids; sicyF: GGTTGTGTTGAGGTTTCGGT and sicyR: TCCGAGCCGAACTAAGTCAA for *Sicydium*).

Effect of increasing number of loci on species discrimination. Our assumption was that individuals of recently diverged species should be more discernible using many loci than using fewer loci. Thus, we calculated p-distance among 10 individuals of *Siniperca*, including five *Sini. chuatsi* and five *Sini. kneri*, using different number of loci to test this hypothesis. Loci with no missing data in all 10 individuals of *Siniperca* were picked using a custom Perl scripts (picktaxagene.pl). The obtained 2,612 loci were then sorted by their average p-distance (distoutlier.pl), so outlier loci with extreme large p-distance could be checked by eye to spot bad data or bad alignment. After removing the bad data, a different number (1, 3, 10, 30, 50, 70, 90, 100, 200, 300, 400, 500, 600, 700, 800, 900, 1000, 2000) of loci were randomly picked and concatenated (samplegene.pl) for calculating p-distance among individuals (gapdis.pl). The sampling at each level of different number of loci was repeated two hundred times. The p-distance among individuals vs. the number of loci used was drawn with GraphPad Prism 5 (San Diego, California). To check the effect of increasing number of loci on species discrimination, the "all species barcodes" criteria was applied, that is queries was considered successfully identified when they were followed by all conspecifics according to their barcode[44]. Custom Perl script was used to calculate the rate of successful identification for 200 replicates at each level of number of loci used (ID_correct_rate.pl). Among individual p-distance and rate of successful identification also were calculated for the *Sicydium*. Sequences of COI gene also were used to calculate p-distance between individuals from the same species and from different species to compare with the results of nuclear genes. Spider[28] was used to optimize barcoding distance threshold and to identify species using COI sequences as suggested by Collins and Cruickshank[27]. The final number of loci recommended for DNA barcoding was chosen based on the effect of increasing number of loci on the success rate of species discrimination.

Estimating species divergence and gene flow in the empirical data. Gene flow and differentiation time of *Sini. chuatsi* and *Sini. kneri* was estimated using IMa2 program with 200 loci[62]. The MCMC was run for 10 million generations with sample recorded every hundred generations. The number of chains was set to 20. The running parameters were set as -q2, -m1, -t3, -b 10000000, d100, -hn20 and -s123. An additional run was performed with the same parameter but different seeds –s111. These two run showed decent mixing, and similar results, so we combined results from the two runs. Similar runs were done for the two species of *Sicydium*. The genetic differentiation between the two species of *Siniperca* and the two species of *Sicydium* also was estimated using Structure 2.3.4[63]. Three iterations for 100,000 generations (using a 100,000 burnin) were run for each value of K (number of population clusters) ranging from 1 to 3. To identify the number of population clusters that captures the major structure in the data, Structure Harvester[64] was used to calculate the peak value for delta K[65].

Simulating sister species sequence data with different divergence times and gene flow. We simulated two diverging species with various splitting time and migration rates to explore the effect of changing these two factors on species discrimination over a broader range of parameter space. According to the IMa2 results of the empirical data, the splitting time was set as 1,000, 10,000, 100,000, and 700,000 generations. The migration rate was set as 0, 0.000001, 0.00001, and 0.0001 per generation. The simulation with 1,000 generations splitting time was combined with only 0 migration rate, because the two simulated species were already indistinguishable under 1,000 generations splitting time even when there was no gene flow in the simulation. The simulations with 10,000, 100,000, and 700,000 generation splitting time were combined with all four migration rates. Fastsimcoal2[66,67] was used to generate the simulated data. Twenty thousand replicates were simulated for each scenario. The effective population size used for simulation was 20,000 in the ancestor species and the two descendant species. The mutation rate was set to 2×10^{-8}. Five sequences were sampled from each simulated species. The simulated data were used to calculate p-distance among individuals of the same and different species. Species identification success rate applying "all species barcodes" criteria was calculated as described above. Identification success rate using different number (1, 3, 5, 10, 30, 50, 70, 90, 100, 200, 300, 400, 500, 600, 700, 800, 900, 1000) of simulated loci was plotted against species splitting time and migration rate using R[68].

A three-step multilocus DNA barcoding pipeline. It is straightforward to use distance based methods to reveal divergence of two sister species in the empirical and simulated data. But for more than two species, distance based species identification becomes more complicated. Firstly, an arbitrary barcoding threshold is needed to judge whether the query is one of the species represented in the database or is a new and distinct species. Secondly, the shortest distance does not necessarily guarantee a sister species relationship, because sister species with long branches might be less similar to the query species than a non-sister species with a short branch. To avoid these risks, we propose a three-step DNA barcoding method (Fig. 6).

In the first step, p-distances between the query and all sequences in the database are calculated. The sequences that are similar to the query are kept for subsequent analyses (p-distance.pl). This is a fast screening process to retrieve all sequences from potential conspecifics or sister species. Because the closest sequence might not be from a conspecifics or sister species, sequences from up to four species are kept. In the second step, a species tree is reconstructed using the sequences from the first step to identify potential conspecifics or sister species of the query using RAxML version 8[69] and ASTRAL 4.10.6[70–72]. Individual gene trees are inferred using RAxML with the GTRGAMMA model, and then a species tree is recovered from those gene trees using ASTRAL. The potential conspecifics or sister species to the query are then chosen based on the phylogenetic relationship depictured in the species tree. In the third step, species delimitation is done using a Bayes factor delimitation approach, BFD*[73]. Single nucleotides polymorphism (SNP) data are retrieved from the sequencing reads of the species chosen in

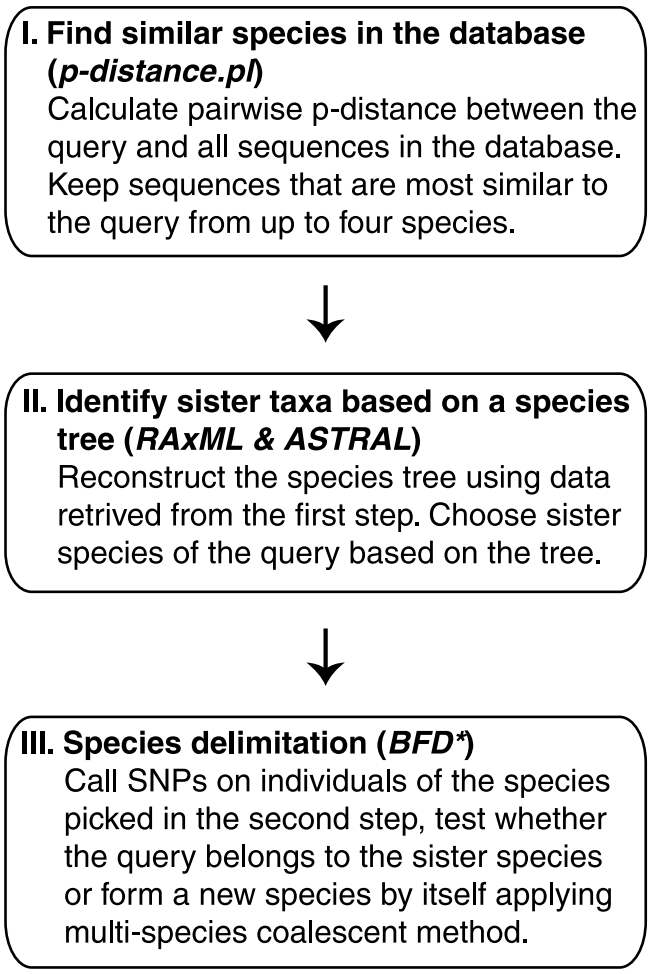

Figure 6. A three-step multilocus DNA barcoding pipeline.

step two and used for the BFD* analysis. A path sampling with 48 steps was conducted to estimate the marginal likelihood with a Markov chain Monte Carlo (MCMC) chain length of 200,000 and a pre-burnin of 50,000 following the recommended settings in BFD*[73]. The strength of support for compared hypotheses was evaluated from Bayes factor scale, 2ln(BF) using the framework of Kass and Raftery[74]. The BF scale is as follows: $0 < 2\ln(BF) < 2$ is not worth more than a bare mention, $2 < 2\ln(BF) > 6$ means positive evidence, $6 < 2\ln(BF) < 10$ represents strong support, and $2\ln(BF) > 10$ represents decisive support. If the result of BFD* analysis does not support two separate species, the query will be assigned to the "sister species"; otherwise, the query will be considered as a new species with its sequences add to the database and further study on its species status will be recommended.

The final set of selected markers was used for testing the above-described three-step multilocus DNA barcoding in the sinipercids, including 26 individuals of seven species. An individual of *Sini. kneri* or *Sini. chuatsi* was randomly chosen as unknown query that needs to be identified. The sequences of the unknown specimens and all other sequences in the database were aligned using Clustal Omega v1.1.1[75]. Custom Perl scripts, concatnexus.pl and gapdis.pl were used to concatenate the sequences of individual loci, to calculate their p-distance between the query and the sample in the database, and sorted them by the p-distance to find all individuals that are close to the query sample.

Testing effect of missing data in the database or in the query on the success rate of species identification. To test if our method could identify new species when the sequences of conspecifics are not in the database, all samples of *Sini. kneri* were removed from the database except that one random selected *Sini. kneri* individual was left as query. To access the effect of missing data in the query sample, one *Sini. kneri* was selected as an unknown sample, and 20 percent, 30 percent, and 50 percent of its loci were excluded, then the data were used for multilocus DNA barcoding analysis.

Data availability. The raw sequence reads are available in NCBI repository (accession: PRJNA373944 and PRJNA355377). The sequences alignments and Perl scripts can be found in Supplementary Information.

References

1. Hassold, S. *et al.* DNA Barcoding of Malagasy Rosewoods: Towards a Molecular Identification of CITES-Listed Dalbergia Species. *PLoS One* **11**, e0157881, https://doi.org/10.1371/journal.pone.0157881 (2016).
2. Candek, K. & Kuntner, M. DNA barcoding gap: reliable species identification over morphological and geographical scales. *Mol Ecol Resour* **15**, 268–277, https://doi.org/10.1111/1755-0998.12304 (2015).
3. Sutou, M., Kato, T. & Ito, M. Recent discoveries of armyworms in Japan and their species identification using DNA barcoding. *Mol Ecol Resour* **11**, 992–1001, https://doi.org/10.1111/j.1755-0998.2011.03040.x (2011).
4. Witt, J. D., Threloff, D. L. & Hebert, P. D. DNA barcoding reveals extraordinary cryptic diversity in an amphipod genus: implications for desert spring conservation. *Mol Ecol* **15**, 3073–3082, https://doi.org/10.1111/j.1365-294X.2006.02999.x (2006).
5. Kadarusman *et al.* Cryptic diversity in Indo-Australian rainbowfishes revealed by DNA barcoding: implications for conservation in a biodiversity hotspot candidate. *PLoS One* **7**, e40627, https://doi.org/10.1371/journal.pone.0040627 (2012).
6. Ghahramanzadeh, R. *et al.* Efficient distinction of invasive aquatic plant species from non-invasive related species using DNA barcoding. *Mol Ecol Resour* **13**, 21–31, https://doi.org/10.1111/1755-0998.12020 (2013).
7. Marescaux, J. & Van Doninck, K. Using DNA barcoding to differentiate invasive Dreissena species (Mollusca, Bivalvia). *Zookeys*, 235–244, https://doi.org/10.3897/zookeys.365.5905 (2013).
8. Saunders, G. W. Routine DNA barcoding of Canadian Gracilariales (Rhodophyta) reveals the invasive species Gracilaria vermiculophylla in British Columbia. *Mol Ecol Resour* **9**(Suppl s1), 140–150, https://doi.org/10.1111/j.1755-0998.2009.02639.x (2009).
9. Hartvig, I., Czako, M., Kjaer, E. D., Nielsen, L. R. & Theilade, I. The Use of DNA Barcoding in Identification and Conservation of Rosewood (Dalbergia spp.). *PLoS One* **10**, e0138231, https://doi.org/10.1371/journal.pone.0138231 (2015).
10. Nevill, P. G., Wallace, M. J., Miller, J. T. & Krauss, S. L. DNA barcoding for conservation, seed banking and ecological restoration of Acacia in the Midwest of Western Australia. *Mol Ecol Resour* **13**, 1033–1042, https://doi.org/10.1111/1755-0998.12060 (2013).
11. Shapcott, A. *et al.* Mapping biodiversity and setting conservation priorities for SE Queensland's rainforests using DNA barcoding. *PLoS One* **10**, e0122164, https://doi.org/10.1371/journal.pone.0122164 (2015).
12. Tanzler, R., Sagata, K., Surbakti, S., Balke, M. & Riedel, A. DNA barcoding for community ecology–how to tackle a hyperdiverse, mostly undescribed Melanesian fauna. *PLoS One* **7**, e28832, https://doi.org/10.1371/journal.pone.0028832 (2012).
13. Hebert, P. D., Cywinska, A., Ball, S. L. & deWaard, J. R. Biological identifications through DNA barcodes. *Proc Biol Sci* **270**, 313–321, https://doi.org/10.1098/rspb.2002.2218 (2003).
14. Vences, M., Thomas, M., Bonett, R. M. & Vieites, D. R. Deciphering amphibian diversity through DNA barcoding: chances and challenges. *Philos Trans R Soc Lond B Biol Sci* **360**, 1859–1868, https://doi.org/10.1098/rstb.2005.1717 (2005).
15. Smith, M. A., Fisher, B. L. & Hebert, P. D. DNA barcoding for effective biodiversity assessment of a hyperdiverse arthropod group: the ants of Madagascar. *Philos Trans R Soc Lond B Biol Sci* **360**, 1825–1834, https://doi.org/10.1098/rstb.2005.1714 (2005).
16. Ward, R. D., Zemlak, T. S., Innes, B. H., Last, P. R. & Hebert, P. D. DNA barcoding Australia's fish species. *Philos Trans R Soc Lond B Biol Sci* **360**, 1847–1857, https://doi.org/10.1098/rstb.2005.1716 (2005).
17. Hawlitschek, O. *et al.* Comprehensive DNA barcoding of the herpetofauna of Germany. *Mol Ecol Resour* **16**, 242–253, https://doi.org/10.1111/1755-0998.12416 (2016).
18. Li, J. *et al.* DNA barcoding of Murinae (Rodentia: Muridae) and Arvicolinae (Rodentia: Cricetidae) distributed in China. *Mol Ecol Resour* **15**, 153–167, https://doi.org/10.1111/1755-0998.12279 (2015).
19. Collins, R. A. *et al.* Barcoding and border biosecurity: identifying cyprinid fishes in the aquarium trade. *PLoS One* **7**, e28381, https://doi.org/10.1371/journal.pone.0028381 (2012).
20. Decru, E. *et al.* Taxonomic challenges in freshwater fishes: a mismatch between morphology and DNA barcoding in fish of the north-eastern part of the Congo basin. *Mol Ecol Resour* **16**, 342–352, https://doi.org/10.1111/1755-0998.12445 (2016).
21. Chan, A. *et al.* DNA barcoding: complementing morphological identification of mosquito species in Singapore. *Parasite Vector* **7**, 569, https://doi.org/10.1186/s13071-014-0569-4 (2014).
22. van Velzen, R., Weitschek, E., Felici, G. & Bakker, F. T. DNA barcoding of recently diverged species: relative performance of matching methods. *PLoS One* **7**, e30490, https://doi.org/10.1371/journal.pone.0030490 (2012).
23. Spasojevic, T., Kropf, C., Nentwig, W. & Lasut, L. Combining morphology, DNA sequences, and morphometrics: revising closely related species in the orb-weaving spider genus Araniella (Araneae, Araneidae). *Zootaxa* **4111**, 448–470, https://doi.org/10.11646/zootaxa.4111.4.6 (2016).
24. Aliabadian, M. *et al.* DNA barcoding of Dutch birds. *Zookeys*, 25–48, https://doi.org/10.3897/zookeys.365.6287 (2013).
25. Mabragana, E. *et al.* DNA barcoding identifies Argentine fishes from marine and brackish waters. *PLoS One* **6**, e28655, https://doi.org/10.1371/journal.pone.0028655 (2011).
26. Dowton, M., Meiklejohn, K., Cameron, S. L. & Wallman, J. A preliminary framework for DNA barcoding, incorporating the multispecies coalescent. *Syst Biol* **63**, 639–644, https://doi.org/10.1093/sysbio/syu028 (2014).
27. Collins, R. A. & Cruickshank, R. H. Known knowns, known unknowns, unknown unknowns and unknown knowns in DNA barcoding: a comment on Dowton *et al.* *Syst Biol* **63**, 1005–1009, https://doi.org/10.1093/sysbio/syu060 (2014).
28. Brown, S. D. *et al.* Spider: an R package for the analysis of species identity and evolution, with particular reference to DNA barcoding. *Mol Ecol Resour* **12**, 562–565, https://doi.org/10.1111/j.1755-0998.2011.03108.x (2012).
29. Puillandre, N., Lambert, A., Brouillet, S. & Achaz, G. ABGD, Automatic Barcode Gap Discovery for primary species delimitation. *Mol Ecol* **21**, 1864–1877, https://doi.org/10.1111/j.1365-294X.2011.05239.x (2012).
30. Virgilio, M., Jordaens, K., Breman, F. C., Backeljau, T. & De Meyer, M. Identifying insects with incomplete DNA barcode libraries, African fruit flies (Diptera: Tephritidae) as a test case. *PLoS One* **7**, e31581, https://doi.org/10.1371/journal.pone.0031581 (2012).
31. Sonet, G. *et al.* Adhoc: an R package to calculate ad hoc distance thresholds for DNA barcoding identification. *Zookeys*, 329–336, https://doi.org/10.3897/zookeys.365.6034 (2013).
32. Li, C., Riethoven, J. J. & Naylor, G. J. EvolMarkers: a database for mining exon and intron markers for evolution, ecology and conservation studies. *Mol Ecol Resour* **12**, 967–971, https://doi.org/10.1111/j.1755-0998.2012.03167.x (2012).
33. Bi, K. *et al.* Transcriptome-based exon capture enables highly cost-effective comparative genomic data collection at moderate evolutionary scales. *BMC Genomics* **13**, 403, https://doi.org/10.1186/1471-2164-13-403 (2012).
34. Hedtke, S. M., Morgan, M. J., Cannatella, D. C. & Hillis, D. M. Targeted enrichment: maximizing orthologous gene comparisons across deep evolutionary time. *PLoS One* **8**, e67908, https://doi.org/10.1371/journal.pone.0067908 (2013).
35. Li, C., Hofreiter, M., Straube, N., Corrigan, S. & Naylor, G. J. Capturing protein-coding genes across highly divergent species. *Biotechniques* **54**, 321–326, https://doi.org/10.2144/000114039 (2013).
36. Jiang, J. *et al.* Gene markers for exon capture and phylogenomics in ray-finned fishes. *bioRxiv*. https://doi.org/10.1101/180786 (2017).
37. Liu, H. & Chen, Y. Phylogeny of the sinipercine fishes with some taxonomic notes. *Zool Res* **15**, 1–12 (1994).
38. Nelson, J. S. *Fishes of the world*. 4th edn, (John Wiley and Sons, Inc., 2006).
39. Zhou, C., Yang, Q. & Cai, D. On the classification and distribution of the sinipercinae fishes (family Serranidae). *Zool Res* **9**, 113–125 (1988).
40. Li, S. Geographical distribution of the Sinipercine fishes. *Chinese J Zool* **26**, 40–44 (1991).
41. Zhao, J., Li, C., Zhao, L., Wang, W. & Cao, Y. Mitochondrial diversity and phylogeography of the Chinese perch, Siniperca chuatsi (Perciformes: Sinipercidae). *Mol Phylogenet Evol* **49**, 399–404 (2008).

42. Bussing, W. A. *Sicydium adelum*, a new species of gobiid fish (Pisces: Gobiidae) from Atlantic slope streams of Costa Rica. *Rev Biol Trop* **44**, 819–825 (1996).
43. Chabarria, R. E. *Evolution of the genus Sicydium* (*Gobiidae: Sicydiinae*). PhD thesis, Texas A&M University – Corpus Christi (2015).
44. Meier, R., Shiyang, K., Vaidya, G. & Ng, P. K. DNA barcoding and taxonomy in Diptera: a tale of high intraspecific variability and low identification success. *Syst Biol* **55**, 715–728, https://doi.org/10.1080/10635150600969864 (2006).
45. Kumar, S. & Subramanian, S. Mutation rates in mammalian genomes. *Proc Natl Acad Sci USA* **99**, 803–808, https://doi.org/10.1073/pnas.022629899 (2002).
46. Lee, J. Y. & Edwards, S. V. Divergence across Australia's Carpentarian barrier: statistical phylogeography of the red-backed fairy wren (Malurus melanocephalus). *Evolution* **62**, 3117–3134 (2008).
47. Funk, D. J. & Omland, K. E. Species-Level Paraphyly And Polyphyly: Frequency, Causes, and Consequences, with Insights from Animal Mitochondrial DNA. *Annu Rev Ecol Evol Syst* **34**, 397–423 (2003).
48. Ross, H. A. The incidence of species-level paraphyly in animals: a re-assessment. *Mol Phylogenet Evol* **76**, 10–17, https://doi.org/10.1016/j.ympev.2014.02.021 (2014).
49. Shen, X. X., Liang, D., Wen, J. Z. & Zhang, P. Multiple genome alignments facilitate development of NPCL markers: a case study of tetrapod phylogeny focusing on the position of turtles. *Mol Biol Evol* **28**, 3237–3252, https://doi.org/10.1093/molbev/msr148 (2011).
50. McCormack, J. E. *et al*. A phylogeny of birds based on over 1,500 loci collected by target enrichment and high-throughput sequencing. *PLoS One* **8**, e54848, https://doi.org/10.1371/journal.pone.0054848 (2013).
51. Yuan, H. *et al*. Target gene enrichment in the cyclophyllidean cestodes, the most diverse group of tapeworms. *Mol Ecol Resour*. https://doi.org/10.1111/1755-0998.12532 (2016).
52. Young, A. D. *et al*. Anchored enrichment dataset for true flies (order Diptera) reveals insights into the phylogeny of flower flies (family Syrphidae). *BMC Evol Biol* **16**, 143, https://doi.org/10.1186/s12862-016-0714-0 (2016).
53. Schmickl, R. *et al*. Phylogenetic marker development for target enrichment from transcriptome and genome skim data: the pipeline and its application in southern African Oxalis (Oxalidaceae). *Mol Ecol Resour* **16**, 1124–1135, https://doi.org/10.1111/1755-0998.12487 (2016).
54. Hugall, A. F., O'Hara, T. D., Hunjan, S., Nilsen, R. & Moussalli, A. An Exon-Capture System for the Entire Class Ophiuroidea. *Mol Biol Evol* **33**, 281–294, https://doi.org/10.1093/molbev/msv216 (2016).
55. Blaimer, B. B., Lloyd, M. W., Guillory, W. X. & Brady, S. G. Sequence Capture and Phylogenetic Utility of Genomic Ultraconserved Elements Obtained from Pinned Insect Specimens. *PLoS One* **11**, e0161531, https://doi.org/10.1371/journal.pone.0161531 (2016).
56. Coissac, E., Hollingsworth, P. M., Lavergne, S. & Taberlet, P. From barcodes to genomes: extending the concept of DNA barcoding. *Mol Ecol* **25**, 1423–1428, https://doi.org/10.1111/mec.13549 (2016).
57. Li, X. *et al*. Plant DNA barcoding: from gene to genome. *Biol Rev Camb Philos Soc* **90**, 157–166, https://doi.org/10.1111/brv.12104 (2015).
58. Heled, J. & Drummond, A. J. Bayesian inference of species trees from multilocus data. *Mol Biol Evol* **27**, 570–580, https://doi.org/10.1093/molbev/msp274 (2010).
59. Rannala, B. & Yang, Z. Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci. *Genetics* **164**, 1645–1656 (2003).
60. Yang, Z. & Rannala, B. Bayesian species delimitation using multilocus sequence data. *Proc Natl Acad Sci USA* **107**, 9264–9269, https://doi.org/10.1073/pnas.0913022107 (2010).
61. Song, S., Zhao, J. & Li, C. Species delimitation and phylogenetic reconstruction of the sinipercids (Perciformes: Sinipercidae) based on target enrichment of thousands of nuclear coding sequences. *Mol Phylogenet Evol* **111**, 44–55, https://doi.org/10.1016/j.ympev.2017.03.014 (2017).
62. Hey, J. Isolation with migration models for more than two populations. *Mol Biol Evol* **27**, 905–920, https://doi.org/10.1093/molbev/msp296 (2010).
63. Pritchard, J. K., Stephens, M. & Donnelly, P. Inference of population structure using multilocus genotype data. *Genetics* **155**, 945–959 (2000).
64. Earl, D. A. & vonHoldt, B. M. STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. *Conserv Genet Resour* **4**, 359–361 (2012).
65. Evanno, G., Regnaut, S. & Goudet, J. Detecting the number of clusters of individuals using the softwareSTRUCTURE: a simulation study. *Mole Ecol* **14**, 2611–2620, https://doi.org/10.1111/j.1365-294X.2005.02553.x (2005).
66. Excoffier, L. & Foll, M. fastsimcoal: a continuous-time coalescent simulator of genomic diversity under arbitrarily complex evolutionary scenarios. *Bioinformatics* **27**, 1332–1334, https://doi.org/10.1093/bioinformatics/btr124 (2011).
67. Excoffier, L., Dupanloup, I., Huerta-Sanchez, E., Sousa, V. C. & Foll, M. Robust demographic inference from genomic and SNP data. *PLoS Genet* **9**, e1003905, https://doi.org/10.1371/journal.pgen.1003905 (2013).
68. R: A language and environment for statistical computing (R Foundation for Statistical Computing, Vienna, Austria, 2015).
69. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313, https://doi.org/10.1093/bioinformatics/btu033 (2014).
70. Mirarab, S., Bayzid, M. S. & Warnow, T. Evaluating Summary Methods for Multilocus Species Tree Estimation in the Presence of Incomplete Lineage Sorting. *Syst Biol* **65**, 366–380, https://doi.org/10.1093/sysbio/syu063 (2016).
71. Mirarab, S. *et al*. ASTRAL: genome-scale coalescent-based species tree estimation. *Bioinformatics* **30**, i541–548, https://doi.org/10.1093/bioinformatics/btu462 (2014).
72. Mirarab, S. & Warnow, T. ASTRAL-II: coalescent-based species tree estimation with many hundreds of taxa and thousands of genes. *Bioinformatics* **31**, i44–52, https://doi.org/10.1093/bioinformatics/btv234 (2015).
73. Leaché, A. D., Fujita, M. K., Minin, V. N. & Bouckaert, R. R. Species delimitation using genome-wide SNP data. *Syst Biol* **63**, 534–542, https://doi.org/10.1093/sysbio/syu018 (2014).
74. Kass, R. E. & Raftery, A. E. Bayes factors. *J Am Stat Assoc* **90**, 773–795 (1995).
75. Sievers, F. *et al*. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol Syst Biol* **7**, 539, https://doi.org/10.1038/msb.2011.75 (2011).

Acknowledgements

This work was supported by Shanghai Pujiang Program, Shanghai Universities First-class Disciplines Project of Fisheries, Program for Professor of Special Appointment (Eastern Scholar) at Shanghai Institutions of Higher Learning to C. Li. The authors would like to thank Shanghai Oceanus Supercomputing Center (SOSC) for providing computational resources.

Author Contributions

C.L. conceived the research project. J.L. conducted the experiments, data analysis and simulation. S.S provided gene capture data of the sinipercids. J.J., L.T. and R.C. read the manuscript and advised on method development. J.L., G.J.P.N., L.T. and C.L. wrote the manuscript.

Additional Information

Supplementary information accompanies this paper at https://doi.org/10.1038/s41598-017-16920-2.

Competing Interests: The authors declare that they have no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.