Molecular-based rapid inventories of sympatric diversity: A comparison of DNA barcode clustering methods applied to geography-based vs clade-based sampling of amphibians

Andrea Paz[1,*] and Andrew J Crawford[1,2]

[1]Department of Biological Sciences, Universidad de los Andes, A.A. 4976, Bogotá, Colombia

[2]Smithsonian Tropical Research Institute, Apartado 0843-03092, Panamá, Republic of Panama

*Corresponding author (Fax, +57-1-332-4069; Email, paz.andreita@gmail.com)

Molecular markers offer a universal source of data for quantifying biodiversity. DNA barcoding uses a standardized genetic marker and a curated reference database to identify known species and to reveal cryptic diversity within well-sampled clades. Rapid biological inventories, e.g. rapid assessment programs (RAPs), unlike most barcoding campaigns, are focused on particular geographic localities rather than on clades. Because of the potentially sparse phylogenetic sampling, the addition of DNA barcoding to RAPs may present a greater challenge for the identification of named species or for revealing cryptic diversity. In this article we evaluate the use of DNA barcoding for quantifying lineage diversity within a single sampling site as compared to clade-based sampling, and present examples from amphibians. We compared algorithms for identifying DNA barcode clusters (e.g. species, cryptic species or Evolutionary Significant Units) using previously published DNA barcode data obtained from geography-based sampling at a site in Central Panama, and from clade-based sampling in Madagascar. We found that clustering algorithms based on genetic distance performed similarly on sympatric as well as clade-based barcode data, while a promising coalescent-based method performed poorly on sympatric data. The various clustering algorithms were also compared in terms of speed and software implementation. Although each method has its shortcomings in certain contexts, we recommend the use of the ABGD method, which not only performs fairly well under either sampling method, but does so in a few seconds and with a user-friendly Web interface.

[Paz A and Crawford AJ 2012 Molecular-based rapid inventories of sympatric diversity: A comparison of DNA barcode clustering methods applied to geography-based vs clade-based sampling of amphibians. *J. Biosci.* **37** 887–896] **DOI** 10.1007/s12038-012-9255-x

1. Introduction

Our knowledge of biodiversity is still incomplete. To date, about 14% of the estimated extant species have been taxonomically described and certain groups such as insects and plants are known only very poorly (Blaxter 2003; May and Harvey 2009; Mora *et al.* 2011). Molecular tools are of increasing interest for the discovery and identification of species (Baker and Palumbi 1994). In cases where the tools of traditional taxonomic identification are difficult to apply, such as for juvenile forms, phenotypically highly plastic species or from fragments of specimens, molecular tools have become invaluable (Floyd *et al.* 2002; Neigel *et al.* 2007; Gonzalez *et al.* 2009; Eaton *et al.* 2010; Lumbsch and Leavitt 2011).

Over the last decade an initiative has emerged to obtain DNA barcodes for an appreciable fraction of all known eukaryotic species (Hebert *et al.* 2003a, b; Blaxter 2004). A DNA barcode is defined as a small, standardized fragment of DNA used for species identification through comparison

Keywords. ABGD; biodiversity inventory; cluster analysis; cryptic species; cytochrome oxidase subunit I; DNA barcode of life; Fuzzy Identification; GMYC; SAP

Supplementary materials pertaining to this article are available on the *Journal of Biosciences* Website at *http://www.ias.ac.in/jbiosci/nov2012/supp/Paz.pdf*

Published online: 15 October 2012

with a curated reference library of DNA sequences from known species, and has been suggested as a powerful tool for rapid identification of species, whether from whole specimens, tissue fragments or processed natural products (Hebert *et al.* 2003a; Neigel *et al.* 2007). Successful identification, however, is highly dependent on a complete reference database that allows matching of unknown sequences to previously described species (Ross *et al.* 2003; Will and Rubinoff 2004; Hajibabaei *et al.* 2005; Wilson *et al.* 2011).

Given a curated reference library of DNA sequences from known species, an increasing number of analytical methods are available to assign query sequences to taxa. Using the classical genetic similarity approach, assigning a DNA sequence to a species relies on the existence of a difference between intra- versus inter-specific genetic distances known as the barcode gap (Meyer and Paulay 2005). One may use a threshold of genetic distance to assign DNA sequences to species (Hebert *et al.* 2003a), although applying the same cut-off across clades may be problematic (Will and Rubinoff 2004). Using a relative threshold, such as the 10X rule of inter- versus intra-specific divergence (Hebert *et al.* 2004), may result in an underestimation of biological species (Hickerson *et al.* 2006), and may be sensitive to an incomplete reference library (Moritz and Cicero 2004). Newer methods of species identification include coalescent theory-based approaches (Pons *et al.* 2006; Monaghan *et al.* 2009), Bayesian assignment (Munch *et al.* 2008a, b) and fuzzy sets (Zhang *et al.* 2011). We explain and explore these methods below and compare their application to geography- versus clade-based sampling.

In the case of animals, a fragment of the mitochondrial gene, cytochrome c oxidase subunit I (COI), is widely accepted as a standard DNA barcode (Hebert *et al.* 2003b). The use of COI has been applied to a limited sampling amphibians (Smith *et al.* 2008), and yet two difficulties were encountered, i.e. priming sites are highly variable among species and GenBank lacked comparative data for COI relative to the large ribosomal subunit, 16S. As this latter gene offered more universal primers, it was recommended as an additional barcode for amphibians (Vences *et al.* 2005a, b). Improved COI primers are now available (Che *et al.* 2012), however, and the COI reference database for amphibians continues to grow (e.g. Alonso *et al.* 2012; Pinto-Sánchez *et al.* 2012). Given its higher rate of evolution, COI may also perform better than 16S for some amphibians (Xia *et al.* 2012). We therefore support the view that 16S is a complementary marker, not an alternative marker, to COI (e.g. Crawford *et al.* 2010).

Here we present an evaluation of four methods for clustering DNA barcodes in terms of accuracy in assigning unknown specimens to species, computation time and user friendliness, as applied to two contrasting datasets. Evaluations were performed with a dataset representing all species of amphibians of a single site in Panama (Crawford *et al.* 2010) and a dataset of DNA barcodes representing a clade of African mantellid frogs (Vences *et al.* 2005a, b). The former dataset represents complete sampling from a geographic perspective, i.e. covering all known species and various cryptic species from a single 4 km×4 km area, whereas the latter dataset represents intensive sampling from a phylogenetic perspective, i.e. covering various genera and species within a taxonomic family regardless of their geographic origin.

Completely sampled clades are the goal of traditional phylogenetic studies, whereas biodiversity inventories are most commonly focused on single geographic sites (Vonesh *et al.* 2009). Under phylogenetic sampling, a given dataset should contain all taxa, but may actually contain less phylogenetic depth (since members of non-focal clades are excluded). Under geographic sampling, allopatric sister lineages will not be represented in the dataset, and genetic distances between sampled species should be higher (and phylogenetic depth may be greater). As DNA barcoding studies become integrated with biodiversity surveys and rapid inventories (Janzen *et al.* 2009), the number of geography-based datasets will increase. We therefore sought to compare the performance of barcode cluster identification algorithms under these two contrasting sampling designs.

2. Methods

2.1 *Datasets*

Analyses were conducted on two previously published datasets (supplementary table 1). The first dataset represented total amphibian diversity in the G. D. Omar Torrijos H. National Park near El Copé, Coclé, Panama, representing 63 species, 38 genera, 15 families and all three orders of amphibians (Crawford *et al.* 2010). DNA sequence data from two mitochondrial genes, COI (648 bp with no length variation) and 16S (436 aligned bp, excluding gapped sites), were included, and clustering algorithms (see below) were applied only to samples with data for both genes, i.e. a total of 272 samples. The 16S gene fragments were aligned using ClustalX 2.0 (Thompson *et al.* 1997), and the preferred alignment is available from TreeBASE Study ID number S10283. The second dataset contained 52 COI sequences (583 aligned bp) representing 29 species distributed among four genera and one taxonomic family, Mantellidae (with the exception of *Ptychadena mascareniensis*), from across Madagascar (Vences *et al.* 2005a, b). Alternative taxonomy places all Madagascar samples in one family, Ranidae *sensu lato*.

2.2 Species identification

DNA sequence-based identification of an unknown biological sample commonly involves the comparison of a query sample against a library of reference sequences representing as nearly as possible a complete clade, e.g. the comparison of blood or feathers from bird-strikes on aircraft in North America against an extensive database of COI barcodes developed in collaboration with the United States' Federal Aviation Administration (Kerr *et al.* 2007). In studies of lesser-known faunal and floral communities, in contrast, the investigator often needs to simultaneously create the reference database, identify samples and flag potential cryptic species for further study. In the context of biodiversity inventories, therefore, DNA barcoding becomes a two-step process: first, counting the number of genetic clusters or evolutionary lineages and, second, assigning names to clusters, where possible (Crawford *et al.* 2010).

2.3 Classical barcode gap analysis

Assigning DNA sequences to known species is relatively straightforward when the barcode gap is well demarcated. In the ideal scenario, one may apply a threshold, x, of genetic divergence and any pair of lineages with greater than x divergence will be counted as belonging to distinct species (Hebert *et al.* 2004). The twin distributions of intra- versus inter-specific divergence typically overlap, however, and the application of a potentially conservative threshold may fail to identify younger species (Hickerson *et al.* 2006). Initial observations suggest that for diverse clades of rapidly speciating organisms one might expect to find substantial overlap between inter- and intra-specific divergences (Meier *et al.* 2006), whereas for older species and sympatric samples, the gap could be well demarcated (Crawford *et al.* 2010).

2.4 Automatic Barcode Gap Discovery (ABGD)

Automatic Barcode Gap Discovery (ABGD) provides an efficient algorithm that allows one to partition a DNA sequence dataset into clusters of like taxa, i.e. candidate species or 'primary species hypotheses,' according to a range of potential barcode gap thresholds (Puillandre *et al.* 2011). As with 'classical' DNA barcode gap analysis, two DNA sequences will be considered members of different groups if their genetic distance is bigger than a threshold distance. ABGD provides three main advances over traditional analyses, however. Potential threshold values are obtained from the data themselves (not *a priori*), the algorithm provides potential clustering schemes based on a wide range of potential thresholds and, through an iterative procedure, different 'clades' within the same dataset may be assigned different thresholds (Puillandre *et al.* 2011). The ABGD method was implemented using the Web interface at *http://wwwabi.snv.jussieu.fr/public/abgd/abgdweb.html*. Default parameters were chosen for a first analysis using Kimura 2-parameter (K2P) distances that correct for transition rate bias (relative to transversions) in the substitution process (Kimura 1980). The default for the minimum relative gap width was set to different values between 0 and 1. For both the Panama and the Malagasy datasets, the analysis was performed in less than a minute and does not require any specific computational resources since it is implemented through a Web interface.

2.5 Fuzzy-theory-based identification

Most methods of species assignment make binary decisions regarding classification of unknown samples, i.e. the sample belongs or does not belong to taxon X. DNA barcoding uses a single gene (or combination of genes) that constitutes incomplete information to assign an unknown specimen to a species. Species membership, therefore, could be viewed in probabilistic terms, i.e. understood on a continuous scale. Fuzzy set theory allows for a scaled evaluation of membership in which species assignment takes a value between 0 and 1, referred to as a fuzzy membership function (FMF) (Zhang *et al.* 2011). Analogous to a barcode gap analysis, fuzzy membership is evaluated using two parameters, the maximum intra-specific genetic distance (θ_1) and minimum inter-specific genetic distance (θ_2), both calculated as K2P distances directly from each potential species, when possible, and not averaged across the dataset, i.e. parameters are computed using all individuals from the known species against which a query sequence is being compared, along with all members of the nearest neighbouring species in the dataset. When a species is limited to just one sample, however, θ_1 is estimated as an average among species. Thus, the fuzzy method is also able to assign singleton samples to their own group, which helps alleviate the problem of false-positive assignments (Zhang *et al.* 2011). Each sequence in a dataset of size n acts both as part of the reference database of $n-1$ sequences against which each sequence is compared individually, and as a query against all other sequences. For the first dataset the analysis took 2 days, for the second dataset the analysis was performed in about 1 h on a desktop computer 2.60 GHz AMD Athlon 64X2 Dual Core with Windows 7 operating system.

2.6 Statistic Assignment Package (SAP)

This method uses a Bayesian phylogenetic approach to species assignment. For every query sequence a database, first a search is performed using BLAST (Altschul *et al.* 1990) to find homologous DNA sequences in GenBank (Munch *et al.* 2008a, b). Only homologues with a BLAST score equal to or greater than half of the best matching homologue are used, except when fewer than 50 best hits are recovered, in which case less similar sequences are included until a set of 50 potential homologues is obtained. Sequences are then aligned using ClustalW (Thompson *et al.* 1994) and a large set of phylogenetic trees is sampled using Bayesian Markov chain Monte Carlo sampling of the posterior distribution of trees (Rannala and Yang 1996; Ronquist and Huelsenbeck 2003). Lastly, every query sequence is assigned a posterior probability of belonging to nested monophyletic groups corresponding to the taxonomic hierarchy (Munch *et al.* 2008a, b). In the case of the Panama dataset the analysis was performed using default parameters and the program took about 2 weeks to finish. Eleven sequences in particular caused the program to crash. Fortunately, incomplete runs can be restarted without losing prior results. These troublesome sequences were eliminated from the input in order to continue with the analysis (supplementary table 2), resulting in 261 out of the 272 sequences used as query sequences. The analysis was also performed using only COI sequences for the Panama dataset, in which case the analysis took 17 h. For the Malagasy dataset the analysis was performed using 52 sequences and the program took 210 min to finish. All analyses were performed on a Mac OSX 2.66 GHz Intel Core 2 Duo.

2.7 *General Mixed Yule-Coalescent approach (GMYC)*

The evolutionary conversion of intra- into inter-specific genetic variation implies a reduction in gene flow (Mayr 1942), and the longer two populations have been isolated, the longer will be the expected time until their separated lineages coalescence (Hudson 1990; Slatkin 1991). Longer isolation also implies greater opportunities for speciation under a neutral, allopatric divergence, e.g. the Dobzhansky–Muller model of reproductive isolation (Kulathinal and Singh 2008). Thus, conspecific lineages should show a high rate of coalescence relative to a slower rate for heterospecific lineages. The GMYC approach uses a likelihood-based analysis to identify the point of transition between within-population rates of coalescence versus inter-specific rates of lineage coalescence (Pons *et al.* 2006; Monaghan *et al.* 2009). Each transition identifies an independent population genetic entity, and each entity or cluster may correspond potentially to a biological species. The GMYC model is implemented in the SPLITS package for R (available at *http://r-forge.r-project.org/projects/splits*). GMYC can be implemented using either a single threshold (Pons *et al.* 2006) model in which a single point of transition between intra- and interspecific rates of coalescence is estimated, or a multiple threshold model in which the point of transition is allowed to vary across a genealogy (Monaghan *et al.* 2009). Because GMYC uses genealogical information rather than simply genetic distances, this analysis requires as input an ultrametric gene genealogy. For the Panama dataset we used the tree published in Crawford *et al.* (2010). For the Malagasy dataset an ultrametric Bayesian consensus tree was inferred (figure 1) using the program BEAST v.1.6.1 (Drummond and Rambaut 2007). We used a relaxed clock with an uncorrelated lognormal distribution for the variation of substitution rates, with the mean substitution rate fixed to 1 to obtain a tree with time in units of substitutions per site (Drummond *et al.* 2006).

3. Results

3.1 *Species identification: Panama dataset*

For the Panama dataset results of 2-gene barcode gap analysis and GMYC cluster analysis were reported previously (Crawford *et al.* 2010). Traditional barcode gap analysis appeared to be largely unambiguous for DNA sequences from sympatric specimens, especially for the COI gene (figure 1 in Crawford *et al.* 2010), as there were no small heterospecific genetic distances. Very large conspecific distances were interpreted as revealing candidate or cryptic species (Crawford *et al.* 2010). *Silverstoneia nubicola* showed intermediate levels of divergence and was counted in that study as two lineages (i.e. containing one cryptic lineage). In total, 74 lineages were identified among 63 named species. In contrast, the GMYC method identified 112 clusters including 52 represented by only one sample, thus over-splitting the lineages relative to traditional barcoding (Crawford *et al.* 2010).

ABGD applied to the same 2-gene Panama dataset detected 70 species based on an intra-specific threshold of 6% genetic divergence and including recursive evaluation of group splitting. An initial 6% was chosen over smaller values because of the high inter-specific divergence in amphibians (Johns and Avise 1998; Vences *et al.* 2005a). All but one cluster was the same as found with the classical barcode gap approach (Crawford *et al.* 2010). The ABGD analysis lumped all samples of *Silverstoneia nubicola*, whereas Crawford *et al.* (2010) included one candidate species, *S. nubicola* A versus *S. nubicola* B.

The fuzzy identification analysis matched the previous barcode gap analysis in recovering all 47 species with more than one sample using an FMF threshold of 0.90. Membership probabilities ranged from 0.94 to 1. However, if the separation between two mitochondrial lineages of *S. nubicola* is not specified in the input file (i.e. introducing sequences with the same species names), the program groups all *S. nubicola* samples with an FMF value of 0.93 (thus supporting the results from ABGD for this taxon). The 16

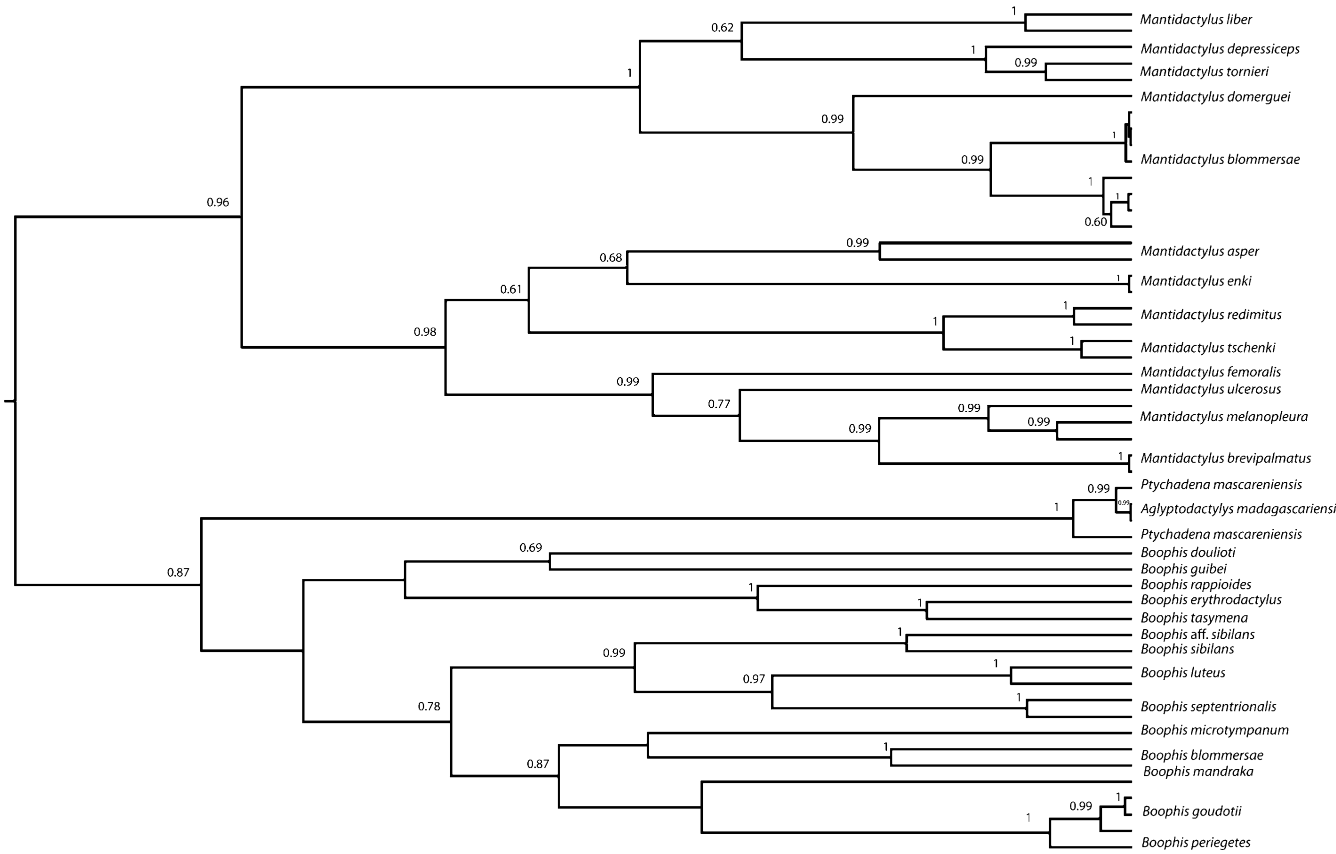

Figure 1. Bayesian phylogeny of the Malagasy dataset. The phylogeny was constructed with 52 COI sequences using the program BEAST v.1.6.2. Numbers represent Bayesian posterior probabilities. See text for details of analysis.

singleton sequences in the dataset were assigned to species with low FMF values (membership probabilities) indicating, correctly, that they were heterospecific from the other samples in the dataset, i.e. low assignment probabilities suggest distinct entities. Twelve singletons were provisionally assigned with membership values between 0 and 0.13 and four were assigned with probabilities ranging from 0.43 to 0.6. The latter four specimens assigned to species with intermediate probabilities were *Agalychnis calcarifer* A, *Lithobates* aff. *warszewitschii*, *Bolitoglossa colonnea* and *B. schizodactyla* (supplementary table 3). The first two specimens correspond to candidate species and were assigned to *A. calcarifer* and *L. watszewitschii* respectively. Again, the low assignment probabilities of the two singleton candidate species reinforces the previous observation based on a barcode gap approach that they are quite distinct entities, at least in terms of their mtDNA (Crawford *et al.* 2010).

Of the 261 combined COI+16S sequences analysed with the SAP package, 100% were successfully assigned to order, 99.6% to family, 98.9% to genus and 93.1% to species. The SAP analysis assigned 243 sequences to species level and identified a total of 59 species. This analysis left 18 samples without a species assignment. These 18 samples corresponded to 8 species, according to the classical barcode gap analysis. Of these unidentified samples, two corresponded to candidate species from Crawford *et al.* (2010), *A. calcarifer* A and *S. nubicola* A, and two samples corresponded to *Dendrobates auratus* and one sample to *Caecilia volcani*. The remaining thirteen unidentified samples corresponded to four species: *Craugastor crassidigitus*, *Hyalinobatrachium colymbiphyllum*, *Cr*. aff. *azueroensis* and *Cochranella euknemos*, for which other samples in the dataset were assigned to species level.

When the SAP analysis was run using only COI sequences, the success rates dropped slightly, relative to the 2-gene analysis. Of the 261 COI sequences, 100% were successfully assigned to order, 97.3% to family, 94.2% to genus and 85.4% to species level. The analysis assigned 223 samples to species level leaving 38 samples without species assignment. These 38 samples correspond to 13 species of which 4 were never assigned to species level: *Oedipina parvipes* and *B. schizodactyla* which where singletons and *Cr.* aff. *azueroensis* and *Cr. fitzingeri* for which no sample was assigned to the species level.

3.2 *Species identification: Malagasy dataset*

ABGD detected 28 lineages among the 29 named species (Vences *et al.* 2005a, b) using an initial 6% intra-specific divergence threshold and recursive splitting of groups. This analysis suggested the existence of cryptic species in two cases. The sequences of *Boophis blommersae* were divided into two different clades as were the sequences belonging to *Mantidactylus asper*. In three other cases, groups previously classified as different species where grouped as single species. *Aglyptodactylus madagascariensis* and *Ptychadena mascareniensis* were lumped into one species. In the genus *Mantidactylus*, *M. depressiceps* and *M. tornieri* were lumped into one entity. In the genus *Boophis* the species *B. goudotii* and *B. periegetes* were grouped together. Thus, the ABGD analysis was congruent with the published taxonomy for 21 out of 29 species, suggesting the existence of cryptic species and a possible over-splitting of three lineages from a purely mtDNA perspective.

In the fuzzy identification analysis, using an FMF threshold of 0.90 to determine successful species assignment, 18 of the 52 sequences were assigned to species (supplementary table 4). Of the remaining 34 sequences, 14 corresponded to singletons that were flagged as potentially unique as evidenced by low FMF values (0 to 0.66). The remaining 20 non-singleton sequences with uncertain assignments (FMF values <0.85) may represent cryptic species, as follows. Consistent with the ABGD analysis, the fuzzy identification results suggested the presence of two groups within both *B. blommersae* and *M. asper*. This analysis also suggested the presence of distinctive groups within *M. brevipalmatus*, *B. goudotii*, *M. tornieri*, *B. luteus*, *M. liber*, *B. septentrionalis*, *M. melanopleura*, *P. mascareniensis* and *M. redimitus*. Two sequences were assigned with high FMF values to taxa different from the *a priori* classification: one sequence of *B. blommersae* was assigned to *M. tschenki* (FMF of 0.94) and one sequence from *B. goudotii* was assigned to *B. microtympanum* (FMF of 0.95). Thus, relative to the published taxonomy, fuzzy identification appeared to over-split some taxa while mis-identifying some COI haplotypes.

The analysis performed with the SAP package resulted in 97.4% of the samples successfully assigned to family, 96.1% to genus and 86.8% to species. Out of the 29 named species entered into the analysis, the SAP package assigned 45 sequences to a total of 24 species. A total of 7 sequences were not assigned to species level. Of these samples, 2 corresponded to *A. madagascariensis*, 3 samples corresponded to singletons entered into the analysis with the names *B. microtympanum*, *B.* aff. *sibilans* and *M. ulcerosus*. The 2 remaining samples corresponded to *P. mascareniensis* and *B. goudotti*, and both species had other samples assigned to species level.

For the Malagasy dataset the GMYC analysis was performed using both the multiple and single threshold methods. However, after performing a likelihood ratio test we discarded the multiple threshold results since this model did not provide a significant improvement in likelihood (P-value$=0.999$). A total of 44 entities, including 6 clusters of more than one sample, were detected in the single threshold analysis. Thus, the analysis resulted in increased splitting of lineages as compared to the 29 named species introduced in the analysis.

4. Discussion

To the extent that allopatric speciation may be more frequent than sympatric speciation (Coyne and Orr 2004), sympatric datasets are less likely to contain sister species, especially for clades with low vagility, such as frogs (Crawford 2003). Sympatric samples may therefore contain more divergent, non-sister species, leading to greater between-species genetic distances. Conspecific samples in sympatry may represent a single population and genetic diversity should be low relative to samples collected from across a species' range. Thus, high between-species distances coupled with minimal conspecific genetic distances should set up an ideal scenario for traditional barcode gap analyses based on genetic distances (Hebert *et al.* 2003a; Meyer and Paulay 2005), and indeed this is what was found previously (Crawford *et al.* 2010). These same conditions would seem to be ideal for other analytical methods, and yet we found some important differences in the performance of each method. When sampling is clade-based, for example, including individuals from within one taxonomic family, it has been argued that the overlap between intra- and inter-specific divergence would be greater and conditions would not be ideal for the application of DNA barcoding (Moritz and Cicero 2004; Meyer and Paulay 2005). In the case of the Malagasy frogs, however, we find that some methods of assignment still perform reasonably well (table 1).

An ideal method for DNA barcode analysis should be able to discriminate successfully between intra- and inter-specific divergence and thus correctly assign query sequences to species. Ideally the method would also be able to detect potential cryptic lineages in a dataset. In the case of single representatives of species in a dataset, i.e. singletons, an ideal method would be able to flag them as unique and not assign them to the nearest (incorrect) species (Lim *et al.* 2011). We evaluated four methods for DNA barcode analysis applied to clade-based sampling and found that their performance varied widely. The GMYC analysis resulted in over-splitting of lineages relative to the other methods, and thus we consider this method as the least conservative regarding species assignment. The SAP method was able to assign 86.8% of samples to species, apparently correctly, and one could regard this as showing better performance than the GMYC method. The FuzzyID and ABGD methods were successful in assigning non-singleton samples to the species

Table 1. A summary of the four DNA barcode identification methods applied to two sampling designs used in this study

Method	Treatment of singletons		Detection of cryptic lineages		Concordance with *a priori* classification	
	Geographic sampling	Clade-based sampling	Geographic sampling	Clade-based sampling	Geographic sampling	Clade-based sampling
SAP	NA*	NA*	NA**	NA**	93.1% of samples assigned to species level.	86.8% of samples assigned to species level.
ABGD	Were correctly left ungrouped.	Two out of fourteen singletons were grouped with incorrect species.	Detected cryptic lineages (as defined in a previous study) in all but one case.	Suggested cryptic lineages within two *a priori* defined species.	Congruent with classical barcode gap analysis in 98% of cases.	Congruent with *a priori* classification in 74% of cases.
GMYC	Many samples were wrongly inferred as singletons.	Some samples were wrongly inferred as singletons.	Suggested the existence of 114 entities in 300 samples. Suggesting cryptic lineages within many of the *a priori* classified species.	Suggested the existence of 44 entities in 52 samples. Suggesting cryptic lineages within most of the *a priori* classified species.	114 entities identified. Over-splitting of groups.	44 entities identified. Slight over-splitting of groups.
FuzzyID	Singletons were correctly flagged as unique.	Singletons were correctly flagged as unique.	Intermediate values of FMF suggested the existence of cryptic lineages congruent with a previous study.	Suggested the existence of many cryptic lineages.	Assignment of 47 specimens with more than one sample congruent with classical barcode gap analysis.	Assignment of 16 out of 52 samples was congruent with previous classification. 2 samples were incongruent. The rest were flagged as unique with low FMF values.

Method	Run time		OS used	Problems	User friendliness	Reference
	Geographic sampling	Clade-based sampling				
SAP	Both genes: 2 weeks. COI only: 17 hours.	210 minutes.	Mac OSX 2.66 GHz Intel Core 2 Duo.	For the geographic-based sampling the program crashed with some samples.	Graphic interface. Easy to use.	(Munch *et al.* 2008a,b)
ABGD	Few seconds	Few seconds.	Web-interface.	Needs *a priori* knowledge of intraspecific divergence in the group.	Web-interface. Easy to use.	(Puillandre *et al.* 2011)
GMYC	Few minutes.	Few minutes.	Mac OSX 2.66 GHz Intel Core 2 Duo.	Requires an ultrametric tree, which takes more time.	Uses R commands but overall easy to use.	(Monaghan *et al.* 2009)
FuzzyID	2 days.	60 minutes.	Windows 7 AMD Athlon™ 64X2 Dual Core.		No Graphic interface and no user manual. Hard to implement.	(Zhang *et al.* 2011)

Methods are compared in terms of speed, performance and user friendliness.

*Singletons are classified as unique in a dataset but the SAP analysis is comparing with an external database where they are not necessarily unique.

**SAP analysis is performed against the GenBank database, and thus can only detect lineages that exist in the database.

level and flagging potential cryptic lineages within two species, *M. asper* and *B. blommersae*, although the FuzzyID method also suggested the existence of additional cryptic lineages (e.g. within *M. brevipalmatus* and within *B. goudotii*). Their relative performance with singletons differed, however. The FuzzyID method was successful in flagging all 14 singletons as unique but the ABGD method wrongly assigned 2 of them to other species: *B. periegetes* and *M. depressiceps* were grouped with *B. goudotii* and *M. tornieri*, respectively. For the geographic-based Panama dataset, ABGD and FuzzyID analyses performed equally well and were congruent with the classical barcode gap analysis used previously (Crawford *et al.* 2010). The GMYC analysis, however, clearly over-split groups, and the SAP analysis was unable to assign some samples to the species level (table 1).

The comparison between Panamanian and Malagasy datasets allowed us to evaluate the performance of four barcode-clustering methods under two distinct sampling designs. Distance-based methods consistently performed better on the geographic-based dataset (Panama) compared to the clade-based sampling (Madagascar), as we predicted. Geography-based sampling within a single site should inherently generate incomplete sampling, which will augment the barcode gap (Moritz and Cicero 2004; Meyer and Paulay 2005). This trend was reversed for the GMYC method, which performed far worse with local, geographic-based sampling. Simulation studies suggest that, in the presence of appreciable migration rates, as the proportion of demes sampled decreases, the GMYC method may overestimate the number of clusters in the dataset (Lohse 2009), although this phenomenon has been little explored empirically. While we still require many more studies of amphibian population structure in this region (but see Crawford 2003; Lampert *et al.* 2003; Robertson *et al.* 2009), our initial findings here support the concerns of Lohse (2009). While the Malagasy dataset also represented a taxonomically incomplete dataset, sampling was at least focused on a single taxonomic family including 28 out of the 191 species of the Mantellidae family (Frost 2009). Including all pairs of sister species could be more problematic for DNA barcoding analyses based on a barcode gap (Meier *et al.* 2006; Moritz and Cicero 2004) and could potentially favour the GMYC approach.

A potential confounding variable in our comparison of the effect of sampling design on DNA barcoding algorithms was the inclusion of a second gene. For the Panama dataset, two mitochondrial genes, COI and 16S, were concatenated and used for most analyses but for the Malagasy dataset only the COI fragment was used. The use of two genes should provide a more accurate genealogical reconstruction, thus potentially favouring a genealogical algorithm such as GMYC, and yet this approach performed worse on the 2-gene Panama dataset. A 2-gene dataset should also increase the number of potential matches in GenBank, thus favouring the SAP approach. When only COI was used for the Panama dataset, 86.8% of the samples were assigned to species level, compared to 93.1% when both COI and 16S were analysed. Thus, a modest increase in performance was achieved adding a second gene to the SAP analysis. Relative to COI, 16S is a slowly evolving gene (Vences *et al.* 2005b), making per-site pairwise distance measures based on a concatenated 16S+COI fragment lower than COI alone. The 1-gene Malagasy dataset might be expected, therefore, to show an advantage when using the distance-based methods such as ABGD and FuzzyID; yet, performance seemed slightly higher for the Panama dataset. Thus, our conclusions can be regarded as conservative concerning the relative success of methods applied to geography-based sampling.

Regardless of the sampling design, the SAP method was unable to assign all samples to species. The SAP approximation is reliant on the existence of a complete reference database. To successfully identify samples, the database must have information from those species. For both datasets, the query samples were already uploaded into GenBank, as we were working with published data. However, with only one sample of one species in the reference database the package cannot reliably assign the unknown sample to a monophyletic group at the species level. In such cases the rest of the orthologues retained from a given BLAST search have lower scores. When the Bayesian phylogeny is reconstructed, the posterior support for a species assignment is low, and thus the analysis can only confidently assign the specimens to higher taxonomic levels. The ABGD method was successful in correctly grouping the unknown samples but requires *a priori* knowledge about the maximum intra-specific divergence of the group under study (P_{max}). ABGD provides different grouping options that depend on the P_{max} value, and the user has to decide which grouping option or options he or she is going to use. In this case we used a value of 6%, which was more in line with a previous study (Crawford *et al.* 2010). However, using a different value of P_{max} will affect the number of clusters inferred for a given dataset. With smaller values of P_{max} more groups will be found, as the variation will be more attributed to inter- rather than intra-specific divergence. In the Panama dataset, using the next lower value of P_{max} (3.6%) did not change the number of groups inferred, but changing to an even lower partition threshold (P_{max}=2.2%) would increase the number of groups from 70 to 74. Using the same value of P_{max} (3.6%) in the Malagasy dataset also results in the same number of groups (28); however, lowing this number still further (P_{max}=2.2%) increased the numbers of groups found by the analysis from 28 to 40. Thus, while ABGD does evaluate clustering under a variety of thresholds, the researcher should have some prior information about divergence levels for a particular group to aid in interpretation of the data.

Systematic comparisons among DNA barcode analysis methods are important for determining the best procedures to rapidly and reliably identify unknown specimens from different type of studies (Reid *et al.* 2011; Boykin *et al.* 2012). The ability to identify species in a fast and reliable way is key for monitoring and controlling illegal animal traffic (Eaton *et al.* 2010; Reid *et al.* 2011), controlling invasive species (Boykin *et al.* 2012), characterizing animal diets (Valentini *et al.* 2008) and studying highly diverse ecosystems such as tropical forests (Kress and Erickson 2008). In some cases combining multiple analytical methods may result in a better identification of unknown samples (Reid *et al.* 2011). Our results suggest that the ABGD method performs well for both sampling designs and does so in just a few seconds. This method is user-friendly and does not have any special computational requirements since it is implemented in a Web-based interface. We thus suggest the use of this method in further DNA barcode studies. As DNA barcoding and automated molecular diversity analyses are more widely applied to rapid assessments of biodiversity inventories, we suggest that the resulting incomplete sampling and exaggerated DNA barcode gap (Moritz and Cicero 2004) will actually facilitate a more accurate count of species diversity within the site, although assigning the correct name to each species may be increasingly difficult.

Acknowledgements

We wish to thank Dr Ramesh K Aggarwal for the kind invitation to participate in this symposium, and the valuable comments of the anonymous reviewer who helped improve this manuscript.

References

Alonso R, Crawford AJ and Bermingham E 2012 Molecular phylogeny of an endemic radiation of Cuban toads (Bufonidae: *Peltophryne*) based on mitochondrial and nuclear genes. *J. Biogeogr.* **39** 434–451

Altschul SF, Gish W, Miller W, Myers EW and Lipman DJ 1990 Basic local alignment search tool. *J. Mol. Biol*. **215** 403–410

Baker CS and Palumbi SR 1994 Which whales are hunted? A molecular genetic approach to monitoring whaling. *Science* **265** 1538–1539

Blaxter M 2003 Molecular systematics: Counting angels with DNA. *Nature* **421** 122–124

Blaxter ML 2004 The promise of a DNA taxonomy. *Philos. Transac. R. Soc. London Ser. B Biol. Sci.* **359** 669–679

Boykin LM, Armstrong KF, Kubatko L and De Barro P 2012 Species delimitation and global biosecurity. *Evol. Bioinformatics* **8** 1–37

Che J, Chen H-M, Yang J-X, Jin J-Q, Jiang KE, Yuan Z-Y, Murphy RW and Zhang Y-P 2012 Universal COI primers for DNA barcoding amphibians. *Mol. Ecol. Resour*. **12** 247–258

Coyne JA and Orr AH 2004 *Speciation* (Sunderland, MA: Sinauer Associates, Inc.)

Crawford AJ 2003 Huge populations and old species of Costa Rican and Panamanian dirt frogs inferred from mitochondrial and nuclear gene sequences. *Mol. Ecol*. **12** 2525–2540

Crawford AJ, Lips KR and Bermingham E 2010 Epidemic disease decimates amphibian abundance, species diversity, and evolutionary history in the highlands of central Panama. *Proc. Natl. Acad. Sci. USA* **107** 13777–13782

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

Drummond AJ and Rambaut A 2007 BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol. Biol*. **7** 214

Eaton M, Meyers G, Kolokotronis S-O, Leslie M, Martin A and Amato G 2010 Barcoding bushmeat: molecular identification of Central African and South American harvested vertebrates. *Conserv. Genet*. **11** 1389–1404

Floyd R, Abebe E, Papert A and Blaxter M 2002 Molecular barcodes for soil nematode identification. *Mol. Ecol.* **11** 839–850

Frost DR 2009 Amphibian species of the world: An online reference version 5.3 (New York: American Museum of Natural History) *http://research.amnh.org/vz/herpetology/amphibia/*

González MA, Baraloto C, Engel J, Mori SA, Pétronelli P, Riéra B, Roger A, Thébaud C and Chave J 2009 Identification of Amazonian trees with DNA barcodes. *PLoS ONE* **4** e7483

Hajibabaei M, deWaard JR, Ivanova NV, Ratnasingham S, Dooh RT, Kirk SL, Mackie PM and Hebert PDN 2005 Critical factors for assembling a high volume of DNA barcodes. *Philos. Transac. R. Soc. London Ser. B Biol. Sci.* **360** 1959–1967

Hebert P, Cywinska A, Ball S and deWaard J 2003a Biological identifications through DNA barcodes. *Philos. Transac. R. Soc. London Ser. B Biol. Sci.* **270** 313–321

Hebert P, Ratnasingham S and deWaard J 2003b Barcoding animal life: cytochrome c oxidase subunit 1 divergences among closely related species. *Philos. Transac. R. Soc. London Ser. B Biol. Sci.* **270** S96–S99

Hebert PDN, Stoeckle MY, Zemlak TS and Francis CM 2004 Identification of birds through DNA barcodes. *PLoS Biol*. **2** e312

Hickerson MJ, Meyer CP and Moritz C 2006 DNA barcoding will often fail to discover new animal species over broad parameter space. *System. Biol*. **55** 729–739

Hudson RR 1990 Gene genealogies and the coalescent process; in *Oxford surveys in evolutionary biology* volume 7 (eds) D Futuyma and J Antonovics (Oxford University Press) pp 1–44

Janzen DH, Hallwachs W, Blandin P, Burns JM, Cadiou J-M, Chacon I, Dapkey T, Deans AR, *et al.* 2009 Integration of DNA barcoding into an ongoing inventory of complex tropical biodiversity. *Mol. Ecol. Resour*. **9** 1–26

Johns GC and Avise JC 1998 A comparative summary of genetic distances in the vertebrates from the mitochondrial cytochome b gene. *Mol. Biol. Evol.* **15** 1481–1490

Kerr KCR, Stoeckle MY, Dove CJ, Weigt LA, Francis CM and Hebert PDN 2007 Comprehensive DNA barcode coverage of North American birds. *Mol. Ecol. Notes* **7** 535–543

Kimura M 1980 A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J. Mol. Evol*. **16** 111–120

Kress WJ and Erickson DL 2008 DNA barcoding - a windfall for tropical biology? *Biotropica* **40** 405–408

Kulathinal R and Singh R 2008 The molecular basis of speciation: from patterns to processes, rules to mechanisms. *J. Genet*. **87** 327–338

Lampert KP, Rand AS, Mueller UG and Ryan MJ 2003 Fine-scale genetic pattern and evidence for sex-biased dispersal in the túngara frog, *Physalaemus pustulosus*. *Mol. Ecol.* **12** 3325–3334

Lim GS, Balke M and Meier R 2011 Determining species boundaries in a world full of rarity: singletons, species delimitation methods. *System. Biol*. **61** 165–169

Lohse K 2009 Can mtDNA barcodes be used to delimit species? A response to Pons *et al.* (2006). *System. Biol*. **58** 439–442

Lumbsch H and Leavitt S 2011 Goodbye morphology? A paradigm shift in the delimitation of species in lichenized fungi. *Fungal Diversity* **50** 59–72

May RM and Harvey PH 2009 Species uncertainties. *Science* **323** 687–687

Mayr E 1942 *Systematics and the origin of species: From the viewpoint of a zoologist* (New York: Columbia University Press)

Meier R, Shiyang K, Vaidya G and Ng P 2006 DNA barcoding and taxonomy in Diptera: a tale of high intraspecific variability and low identification success. *System. Biol*. **55** 715–728

Meyer CP and Paulay G 2005 DNA barcoding: Error rates based on comprehensive sampling. *PLoS Biol*. **3** e422

Monaghan MT, Wild R, Elliot M, Fujisawa T, Balke M, Inward DJG, Lees DC, Ranaivosolo R, Eggleton P, Barraclough TG and Vogler AP 2009 Accelerated species inventory on Madagascar using coalescent-based models of species delineation. *System. Biol*. **58** 298–311

Mora C, Tittensor DP, Adl S, Simpson AGB and Worm B 2011 How many species are there on Earth and in the ocean? *PLoS Biol*. **9** e1001127

Moritz C and Cicero C 2004 DNA barcoding: promise and pitfalls. *PLoS Biol*. **2** e354

Munch K, Boomsma W, Huelsenbeck JP, Willerslev E and Nielsen R 2008a Statistical assignment of DNA sequences using bayesian phylogenetics. *System. Biol*. **57** 750–757

Munch K, Boomsma W, Willerslev E and Nielsen R 2008b Fast phylogenetic DNA barcoding. *Philos. Transac. R. Soc. London Ser. B Biol. Sci.* **363** 3997–4002

Neigel J, Domingo A and Stake J 2007 DNA barcoding as a tool for coral reef conservation. *Coral Reefs* **26** 487–499

Pinto-Sánchez NR, Ibáñez R, Madriñán S, Sanjur OI, Bermingham E and Crawford AJ 2012 The Great American biotic interchange in frogs: Multiple and early colonization of Central America by the South American genus *Pristimantis* (Anura: Craugastoridae). *Mol. Phylogenet. Evol.* **62** 954–972

Pons J, Barraclough TG, Gomez-Zurita J, Cardoso A, Duran DP, Hazell S, Kamoun S, Sumlin WD and Vogler AP 2006 Sequence-based species delimitation for the DNA taxonomy of undescribed insects. *System. Biol*. **55** 595–609

Puillandre N, Lambert A, Brouillet S and Achaz G 2011 ABGD, automatic barcode gap discovery for primary species delimitation. *Mol. Ecol*. **21** 1864–1877

Rannala B and Yang Z 1996 Probability distribution of molecular evolutionary trees: A new method of phylogenetic inference. Journal of Molecular Evolution **43** 304–311

Reid BN, Le M, McCord WP, Iverson JB, Georges A, Bergmann T, Amato G, Desalle R and Naro-Maciel E 2011 Comparing and combining distance-based and character-based approaches for barcoding turtles. *Mol. Ecol. Resour.* **11** 956–967

Robertson JM, Duryea MC and Zamudio KR 2009 Discordant patterns of evolutionary differentiation in two Neotropical treefrogs. *Mol. Ecol*. **18** 1375–1395

Ronquist F and Huelsenbeck JP 2003 MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* **19** 1572–1574

Ross HA, Lento GM, Dalebout ML, Goode M, Ewing G, McLaren P, Rodrigo AG, Lavery S and Baker CS 2003 DNA surveillance: Web-based molecular identification of whales, dolphins, and porpoises. *J. Hered*. **94** 111–114

Slatkin M 1991 Inbreeding coefficients and coalescence times. *Genet. Res.* **58** 167–175

Smith MA, Poyarkov Jr NA and Hebert PDN 2008 CO1 DNA barcoding amphibians: take the chance, meet the challenge. *Mol. Ecol. Res.* **8** 235–246

Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F and Higgins DG 1997 The ClustalX windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res.* **25** 4876–4882

Thompson JD, Higgins DG and Gibson TJ 1994 CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22** 4673–4680

Valentini A, Pompanon F and Taberlet P 2008 DNA barcoding for ecologists. *Trend. Ecol. Evol.* **24** 110–117

Vences M, Thomas M, Bonett RM and Vieites DR 2005a Deciphering amphibian diversity through DNA barcoding: chances and challenges. *Philos. Transac. R. Soc. London Ser. B Biol. Sci.* **360** 1859–1868

Vences M, Thomas M, van der Meijden A, Chiari Y and Vieites DR 2005b Comparative performance of the 16S rRNA gene in DNA barcoding of amphibians. *Front. Zool.* **2** 5

Vonesh JR, Mitchell JC, Howell K and Crawford AJ 2009 Rapid assessments of amphibian diversity; in *Amphibian ecology and conservation: A handbook of techniques* (ed) CK Dodd Jr (Oxford: Oxford University Press) pp 263–280

Will KW and Rubinoff D 2004 Myth of the molecule: DNA barcodes for species cannot replace morphology for identification and classification. *Cladistics* **20** 47–55

Wilson J, Rougerie R, Schonfeld J, Janzen D, Hallwachs W, Hajibabaei M, Kitching I, Haxaire J and Hebert P 2011 When species matches are unavailable are DNA barcodes correctly assigned to higher taxa? An assessment using sphingid moths. *BMC Ecol*. **11** 18

Xia YUN, Gu H-F, Peng RUI, Chen QIN, Zheng Y-C, Murphy RW and Zeng X-M 2012 COI is better than 16S rRNA for DNA barcoding Asiatic salamanders (Amphibia: Caudata: Hynobiidae). *Mol. Ecol. Resour*. **12** 48–56

Zhang AB, Muster C, Liang HB, Zhu CD, Crozier R, Wan P, Feng J and Ward RD 2011 A fuzzy-set-theory-based approach to analyse species membership in DNA barcoding. *Mol. Ecol*. **21** 1848–1863