Methods in Ecology and Evolution

BRITISH ECOLOGICAL SOCIETY

Methods in Ecology and Evolution 2017 doi: 10.1111/2041-210X.12729

Evaluation of short mitochondrial metabarcodes for the identification of Amazonian mammals

Arthur Kocher*[1,2], Benoit de Thoisy[3,4], François Catzeflis[5], Mailis Huguin[3,4], Sophie Valière[6], Lucie Zinger[1], Anne-Laure Bañuls[2] and Jérôme Murienne[1]

[1]*CNRS, University Toulouse III Paul Sabatier, ENFA, UMR5174 EDB (Laboratoire Evolution et Diversité Biologique), Toulouse, France;* [2]*UMR MIVEGEC (IRD 224 – CNRS 5290 – Université de Montpellier), 911 Avenue Agropolis, F34394 Montpellier, France;* [3]*Institut Pasteur de la Guyane, 23 avenue Pasteur, 97300 Cayenne, French Guiana;* [4]*Association Kwata, 16 avenue Pasteur, 97300 Cayenne, French Guiana;* [5]*Institut des Sciences de l'Evolution, Case Courrier 064, CNRS UMR-5554, Université Montpellier-2, Place E. Bataillon, F-34095 Montpellier, France; and* [6]*GeT–PlaGe, Genotoul, INRA Auzeville, 31326 Castanet-Tolosan, France*

Summary

1. DNA barcoding and metabarcoding are increasingly used as alternatives to traditional morphological identifications. For animals, the standard barcode is a c. 658-bp portion of the COI gene, for which reference libraries now cover a large proportion of described mammal species. Unfortunately, because its sequence is too long and does not contain highly conserved primer binding sites, this marker is not adapted for metabarcoding. Although alternative metabarcodes have been developed, their performances are generally seldom assessed.

2. We evaluate the reliability of a short metabarcode located in the mitochondrial 12S ribosomal RNA for the identifications of Amazonian mammals. We (i) constitute a nearly exhaustive reference library for species found in French Guiana, (ii) assess the taxonomic resolution of the marker and validate its use with dipteran blood meal analyses, (iii) assess the conservation of the primer binding sites, and (iv) compare its theoretical performances with that of a newly designed metabarcode located within the standard COI barcode.

3. About 576 specimens representing 164 species were gathered and sequenced. We show that the 12S marker allows remarkably accurate taxonomic assignations despite its very short size, and that primer binding sites are highly conserved, which is important to avoid PCR amplification bias potentially leading to detection failure. Additionally, our results stress that the identifications should only be considered at the generic level when they are based on incomplete reference libraries, even when a stringent similarity cut-off is used. A new short COI metabarcode was designed based on 569 reference sequences of mammals retrieved on BOLD. Our results clearly show that, while both markers provide similar taxonomic resolution, much higher rates of primer mismatches are found with COI.

4. Besides demonstrating the accuracy of the short 12S marker for the identification of Amazonian mammals and providing a reliable molecular reference database, this study emphasizes that the accuracy of taxonomic assignations highly depends on the comprehensiveness of the reference library and that great caution should be taken for interpreting metabarcoding results based on scarce reference libraries. The comparison with a short COI metabarcode also provides novel evidence in support for the use of ribosomal markers in metabarcoding studies.

Key-words: blood meals, COI, French Guiana, mitochondrial 12S rRNA gene, sand fly

Introduction

The accurate identification of species is an essential component in most of the empirical ecological studies. Traditional methods based on morphological features are time consuming and often rely on taxonomic expertise that is increasingly lacking. In addition, morphological identifications may require whole specimens, which can be particularly difficult to obtain for mammals because of the practical, ethical, or legal reasons. DNA-based identification methods have been increasingly used as an efficient alternative over the last decades. Today, one of the most used techniques is DNA barcoding (Hebert *et al.* 2003), which uses the sequence from a short standard fragment of the genome for the taxonomic assignment of a specimen. More recently, high-throughput sequencing has allowed the extension of DNA barcoding for the identification of multiple species in a single sample (Taberlet *et al.* 2012). This approach, referred to as metabarcoding, allows the simultaneous identifications of multiple specimens from a single bulk-DNA extraction (Yu *et al.* 2012; Kocher *et al.* 2016). In addition, it has the great advantage to be applicable on degraded DNA present in the environment such as soil

*Correspondence author. E-mail: arthur.kocher@gmail.com

© 2017 The Authors. Methods in Ecology and Evolution © 2017 British Ecological Society

(Andersen *et al.* 2012) or water (Ficetola *et al.* 2008; Valentini *et al.* 2016). Finally, it constitutes a great tool to study trophic interactions through the analyses of gut content (Coghlan *et al.* 2013) or faeces (Kartzinel *et al.* 2015).

The prerequisites of these methods are the choice of appropriate DNA markers, the design of corresponding PCR primers and the constitution of reliable reference sequences libraries. For DNA barcoding, the Consortium for the Barcode of Life (CBOL, http://www.barcodeoflife.org/) handled these issues, by providing standardized laboratory protocols and curated reference libraries linked to voucher specimens. For animals, the current standard barcode is a c. 658-bp portion of the mitochondrial cytochrome oxidase 1 subunit (COI), and the Barcode of Life Database comprises reference sequences for more than 174 000 animal species to date (BOLD, http://www.boldsystems.org/, accessed in September 2016). Unfortunately, this marker is not the best choice when it comes to metabarcoding (Deagle *et al.* 2014). First, the fragment is too long concerning the limitations of the current sequencing platforms (typically an Illumina Miseq; Illumina, Inc., San Diego, CA, USA). This can be regarded as a rather technical issue that might be overcome in a near future with the rapid improvement of the sequencing technologies. However, the size of the targeted fragment is also critical when dealing with degraded DNA, as typically found in the environment or in biological samples such as faeces or gut content. Second, it is virtually impossible to find perfectly conserved primer binding sites within this coding gene because of high mutation rate at the third codon position (Deagle *et al.* 2014). This may not be a problem for barcoding single specimens, because a few primer-template mismatches will not impede PCR amplification. On the contrary, small variations in the number and position of primer-template mismatches can lead to significant amplification bias or even detection failure when mixtures of DNA are amplified for metabarcoding (Bru, Martin-Laurent & Philippot 2008; Taberlet *et al.* 2012). To find suitable metabarcodes and their associated primers, specific softwares have been developed [notably 'ecoPrimers', (Riaz *et al.* 2011)], that seek to minimize amplification bias while maximizing the divergence between taxa. Most animal metabarcoding markers developed using this approach are located within the mitochondrial ribosomal RNA genes (Riaz *et al.* 2011; Clarke *et al.* 2014; Deagle *et al.* 2014). Indeed, because of the secondary structure of their RNA products, these genes exhibit a mosaic pattern of variation with highly conserved regions (within stems) in which primers can be designed, adjacent to variable regions (within loops) that allow interspecific discrimination. The existence of a standard marker for DNA barcoding has allowed the constitution of a collaborative and taxonomically comprehensive reference library. On the contrary, there is no real consensus on the choice of metabarcoding markers (except for bacteria and fungi), leading to scarce reference libraries (Pompanon & Samadi 2015).

The 12S-V5 marker is a c. 100 base pairs (bp) portion of the mitochondrial 12S ribosomal RNA gene (12S rRNA) (Riaz *et al.* 2011). Based on the sequences available in public databases, it was shown to gather good properties for metabarcoding of vertebrates. However, a comprehensive taxonomic sampling is necessary to precisely assess the taxonomic resolution of a DNA marker (Meyer & Paulay 2005). In this study, we assess the reliability of the 12S-V5 markers for metabarcoding of Amazonian mammals. We (i) constitute a nearly exhaustive reference library for the species found in French Guiana, (ii) assess the variability at the 12S-V5 primer binding sites, (iii) evaluate the taxonomic resolution of the marker and further validate its use with dipteran blood meal analyses, and (iv) compare its theoretical performances with that of a newly designed metabarcode located within the classical COI barcode.

Materials and methods

SAMPLING

French Guiana and its >90% of well-preserved Amazonian rainforest cover has been the study site of intense ecological and taxonomic research (see for instance the research undertaken under the frame of the labex CEBA; http://www.labex-ceba.fr/en/). The mammalian fauna of French Guiana is well characterized, and is representative of a larger part of the northern Amazon region (Lim 2012; Catzeflis 2015). Our aim was to generate a first comprehensive DNA library for the mammals of French Guiana that can be used directly for metabarcoding studies in this region, and that may be further completed for studies in other Amazonian locations. Tissue samples of mammals from French Guiana were gathered from field sampling, museum collections, hunting or road-killed specimens, and biopsies of captured animals (see Supporting Information for details). The taxonomic identifications were based on external and/or craniodental morphology, and confirmed by COI barcoding for the vast majority of the specimens [following classical procedures (Borisenko *et al.* 2008) and the primers C_VF1di/C_VR1LRt1 or LCO1490/HCO2198 recommended by the Barcode of Life Project (www.boldsystems.org)].

REFERENCE LIBRARY

Our aim was to build a reference sequence library based on a marker that was previously developed for the identifications of vertebrates through metabarcoding (12S-V5, Riaz *et al.* 2011). The constitution of reference libraries with the previously designed PCR primers leads to the loss of all the information concerning the primer binding sites. This impedes the possibility to further improve the primers for specific purposes and to predict potential amplification bias (Bru, Martin-Laurent & Philippot 2008). To overcome this issue, we designed a set of primers to amplify a region that contains the complete 12S-V5 fragment including primer binding sites (see Fig. 1). We used blastn 2.2.29+ (Camacho *et al.* 2009) on Genbank (release 197) with the query being a c. 700 bp portion of *Rattus rattus* 12S rRNA (GenBank accession: NC_012374.1) containing the fragment amplified by the 12S-V5 primers and 300 bp flanking each end. We selected all the matching sequences of mammals presenting at least 95% query coverage and kept one sequence per species. The resulting database contained sequences for 1557 mammal species representing 26 orders and was used to design new primers with the ecoPrimers program (Riaz *et al.* 2011). These primers (Mam12S-340-F, 5'-CCACCGCGGTCATACGATT-3'; Mam12S-340-R, 5'-GATGGCGGTATATAGACTG-3') had a maximum of two mismatches with 98·4% of the species represented in the database. They amplify a fragment of 302–350 bp that contains the full

© 2017 The Authors. Methods in Ecology and Evolution © 2017 British Ecological Society,

12S-V5 marker and can be sequenced on an Illumina MiSeq platform (Illumina, San Diego, CA, USA).

DNA AMPLIFICATION AND SEQUENCING

DNA was extracted using the DNeasy Blood and Tissue kit (Qiagen, Valencia, CA, USA). PCR amplification was performed in 25 μL mixtures containing 2 μL of DNA template, 0·2 μL of AmpliTaq Gold® (5 U μL^{-1}, Applied Biosystems, Foster City, CA, USA), 2·5 μL 10X PCR buffer (provided with Amplitaq Gold®, Applied Biosystems), 0·5 μL dNTPs (2·5 mM each, Promega, Madison, WI, USA), 1 μL of each primer (10 μM), 0·25 bovine serum albumin (10 mg mL^{-1}, Promega), 2·5 μL $MgCl_2$ (25 mM, Applied Biosystems) and nuclease-free water (Promega). The PCR mixture was denatured at 95°C (10 min) and followed by 35 cycles of 30 s at 95 °C, 30 s at 50 °C and 30 s at 72 °C, completed at 72 °C (10 min). Tags of eight base pairs with at least five differences between them were added at the 5' end of each primer to enable the sequencing of the multiple PCR products in a single sequencing run (Binladen *et al.* 2007).

PCR products were pooled and sent for library construction and sequencing to the GeT-PlaGe core facilities of Genotoul (Toulouse, France). Samples were diluted in ultrapure water. A volume of 130 μL containing 3 μg of DNA was purified using the HighPrep PCR system (Magbio Genomics, Gaithersburg, MD, USA) and used for library construction with the Illumina NEXTflex PCR-Free DNA sequencing kit following the instructions of the supplier (Bioo Scientific corp., Austin, TX, USA). Purified fragments were end-repaired, A-tailed and ligated to sequencing indexed adapters. The quality of the library was controlled using the Fragment Analyzer (Advanced Analytical, Ames, IA, USA) and quantified by qPCR with the Library Quantification Kit – Illumina Genome Analyzer-SYBR Fast Universal (CliniSciences, Nanterre, France). The library was loaded onto the Illumina MiSeq cartridge according to the manufacturer instructions. The quality of the run was checked internally using PhiX. Quality filtering was performed by the consensus assessment of sequence and variation pipeline. The sequencing data was stored on the NG6 platform (Mariette *et al.* 2012) and all computations were performed on the computer cluster of the Genotoul bioinformatic platform (Toulouse, France).

BIOINFORMATICS

Sequence reads were analysed using the OBITOOLS package (Boyer *et al.* 2016). Pair-end reads were aligned and merged, taking into account the Phred quality scores for consensus construction and alignment score computation. The reads were then assigned to their corresponding sample based on the tagged primer sequences with two mismatches allowed. Low quality reads (alignment scores <50, containing Ns or shorter than 50 bp) were removed. Reads were then dereplicated while keeping the coverage information (number of reads merged). For each sample, the majority sequence was considered as the genuine most abundant sequence in the specimen and kept for the reference library. The script used for these bioinformatic steps are available in the Supporting Information. The library was further completed by 12S rRNA sequences extracted from complete mitogenomes of Xenarthra (Gibb *et al.* 2016) and Chiroptera (F. Botero-Castro, unpublished data).

METABARCODE EVALUATION

Primer-template mismatches were checked by mapping the 12S-V5 primers on the resulting sequences (see Fig. 1; 12S-V5-F: TAGAACAGGCTCCTCTAG; 12S-V5-R: TTAGATACCCCACTATGC), using Geneious 6.0.6 Pro (Biomatters, Auckland, New Zealand). The region corresponding to the 12S-V5 metabarcode was then extracted from each sequence for the following analyses.

Most studies that validate the reliability of DNA barcoding for molecular identifications provide statistics based on K2P genetic distances (Kimura 1980) computed from a multiple sequence alignment of the reference sequences. Uncorrected distances have been judged more appropriate for studying the success of distance-based identification techniques (Srivathsan & Meier 2012). In this study, we used genetic distances as they are computed by the ECOTAG program (included in the OBITOOLS) for taxonomic assignments (uncorrected distances based on pairwise alignments of the sequences) to generate a neighbour-joining tree using the R package 'ape' (Saitou & Nei 1987; Paradis, Claude & Strimmer 2004; Team 2014) and to compute distance statistics using the R package 'spider' (Brown *et al.* 2012).

To assess the reliability of the metabarcode for species identifications, we performed the taxonomic assignment of each specimen using the ECOTAG program with all other sequences as reference library. Resulting assignations were then compared with the genuine identities of the specimens. ECOTAG first searches for the reference sequence(s) showing the highest similarity with the query sequence (primary reference sequence(s); see Fig. 2). Then it looks for all other reference sequences whose similarity with the primary reference sequence(s) is equal or higher than the similarity between the primary reference sequence(s) and the query sequence (secondary reference sequence(s)). Finally, it assigns the query sequence to the most recent common ancestor of the primary and secondary reference sequences. This procedure is similar in essence to the lowest common ancestor algorithms

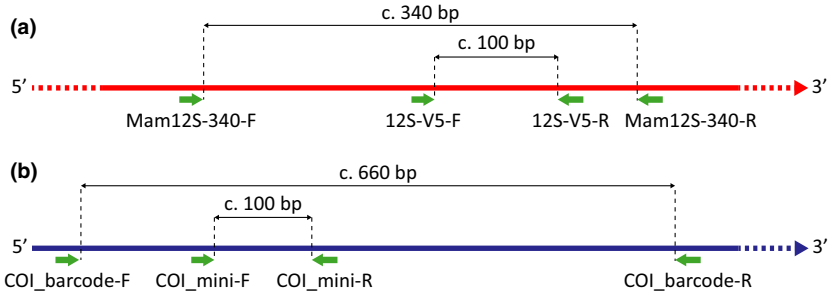

Fig. 1. (a) Relative positions of the Mam12S-340 and 12S-V5 primers binding sites on the 12S mt rRNA gene. PCR amplifications were performed using Mam12S-340 primers to generate a reference library for the 12S-V5 metabarcode while keeping the information concerning the primers binding sites. (b) Relative positions of the standard COI barcode and the newly designed COI_minimam primers on the COI gene. The COI reference library was constituted with full standard barcodes allowing to investigate COI_minimam primers binding sites.

© 2017 The Authors. Methods in Ecology and Evolution © 2017 British Ecological Society,

implemented in the MG-RAST server (Meyer *et al.* 2008) and the MEGAN program (Huson *et al.* 2007) for the assignment of metagenomic reads. It allows to deal with ambiguous identifications, which can arise because several taxa are poorly distinguishable, or that the DNA library does not contain a close reference for the query. Taxonomic assignations were discarded if the closest match exhibited less than 97% similarity. To assess the effect of potential taxonomic gaps in the reference database, we then tested the taxonomic assignations of each sequence after removing all conspecifics. To further validate the applicability of the 12S-V5 marker for mammal species identification in field studies, we analysed blood meals of hematophagous dipteran collected in forest sites in French Guiana (sand flies and mosquitoes, see Supporting Information for details on sampling and laboratory protocols). Indeed, metabarcodes provide good properties (small size and wide taxonomic coverage) for such application, because arthropod blood meals may contain low quantities of degraded DNA from a diverse array of vertebrate species.

COMPARISON WITH COI

Currently, COI is rarely used for metabarcoding because of previously explained reasons (see Introduction). In particular, no satisfying COI metabarcode has been developed for mammals. Therefore, to allow relevant comparison, we designed new PCR primers to amplify a short fragment located within the classical COI barcode. All the sequences of mammal species found in French Guiana were retrieved from BOLD, and a maximum of five sequences per species were kept. PCR primers were designed using ecoPrimers in the same way it was done for the 12S-V5 primers (i.e. 18-bp long, to amplify a fragment of c. 100 bp, and to maximize taxonomic coverage and resolution). The selected primers were compared with their target sites on the reference sequences to compute mismatch statistics. The theoretical amplified fragment was then extracted from references sequences to evaluate its taxonomic resolution in the same way it has been done for the 12S-V5 marker.

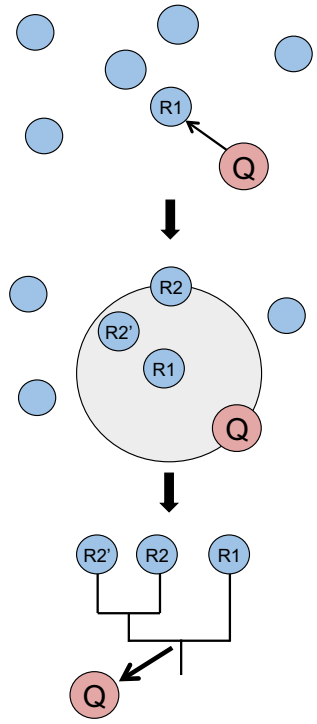

Fig. 2. Schematization of a taxonomic assignation as performed by ecotag: (i) search of the reference sequences (R1s) that have the highest similarity with the query sequence (Q), (ii) search for all other reference sequences (R2s) whose similarity with the primary reference sequences is equal or higher than the similarity between the primary reference sequences and the query sequences, (iii) assignment of the query sequence to the most recent clade containing all R1s and R2s.

Results

REFERENCE LIBRARY

Sequences were obtained for 576 specimens representing 164 species, including *Uroderma* cf. *magnirostrum* (although *U. magnirostrum* has not been officially reported in French Guiana, and no voucher specimen is available, it is not unlikely that this taxon occurs and the samples used in this study have been identified as *U. magnirostrum* using classical COI barcoding). Hence, 82·7% [163/197; (Catzeflis 2015)] of the mammal species recorded in French Guiana were included in the database, whereas 116 over 126 genera and all the 32 families were represented. All the sequences were deposited on Genbank (accessions: KX381203–KX381784). On average, 3·5 specimens per species were sequenced with 118 species (72%) being represented by at least 3 specimens and 28 (17%) represented by only one. The sequence length of the Mam12S-340 fragment ranged from 334 to 345 bp whereas the length of the 12S-V5 fragment ranged from 96 to 103 bp.

METABARCODE EVALUATION

The inspection of the 12S-V5 primer binding sites revealed that the forward primer could be slightly improved by degenerating the 5' end (12S-V5-F': YAGAACAGGCTCCTCTAG). 95·0, 2·8 and 2·2% of the sequences had respectively 0,1, and 2

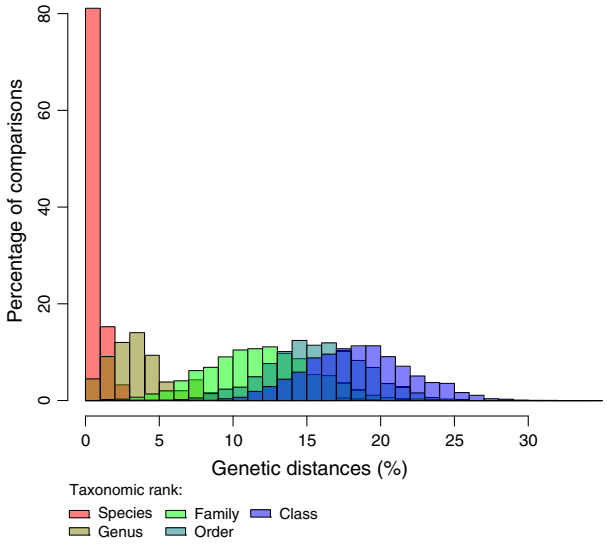

Fig. 3. Distribution of genetic distance at various taxonomic resolutions. Only the distances between specimens belonging to (i) the same taxon at a given rank and to (ii) distinct taxa at the inferior rank are considered.

© 2017 The Authors. Methods in Ecology and Evolution © 2017 British Ecological Society,

mismatches with the forward primer (mean number of mismatches = 0·072; degenerated version), whereas 97·6 and 2·4% had respectively 0 and 1 mismatch with the reverse primers (mean number of mismatches = 0·024). No sequence had mismatches with both primers. No sequence had mismatch towards the 3' end on the forward primer (within the first eight positions). One mismatch was found at the second position from the 3' end of the reverse primer in *Tonatia saurophila* and *Rhynophylla pumilio*, which may hamper amplification for these species.

In the 12S-V5 metabarcode, 34·3% of sites were identical in all sequences and the mean pairwise identity was of 80·2%. The distribution of genetic distances at the specific, generic, familial and ordinal levels shows significant overlap between each consecutive taxonomic rank (Fig. 3). In particular, intraspecific distances range from 0·0 to 5·1% for a mean of 0·5%, while interspecific distances within the same genus range from 0·0 to 19·6% for a mean of 3·5%. The neighbour-joining tree is shown in Fig. 4. Some closely related species were poorly or not distinguishable based on the marker, mostly in Chiroptera and in Felidae: *Pteronotus rubiginosus* and *Pteronotus* cf. sp. 3, *Molossus molossus* and *Molossus rufus*, *Carollia brevicauda* and *Carollia perspicillata*, *Eptesicus furinalis* and *Eptesicus chiriquinus*, *Artibeus planirostris* and *Dermanura gnoma*, *Oecomys rex* and *Oecomys* sp. 1, *Puma concolor* and *Puma yagouaroundi* and *Leopardus wiedii* and *Leopardus pardalis*.

When considering the species represented by more than one specimen (548 sequences), 90·0 and 9·5% of the assignations were made at the species and genus level, respectively (Table 1). 99·6% of these assignations were correct. Only two errors were found: a sequence of *L. wiedii* was assigned to *L. pardalis* and a sequence of *D. gnoma* was assigned to the genus *Artibeus*. Most of the assignations made at the genus level were found in Chiroptera (genera *Carollia*, *Eptesicus*, *Molossus*, and *Pteronotus*). Three sequences (two *Glyphonycteris sylvestris* and one *Makalata didelphoides*) were not identified because of the absence of a close match in the reference database (>97% similarity). When all conspecifics were removed from the references before taxonomic assignment, 65·2% of the sequences

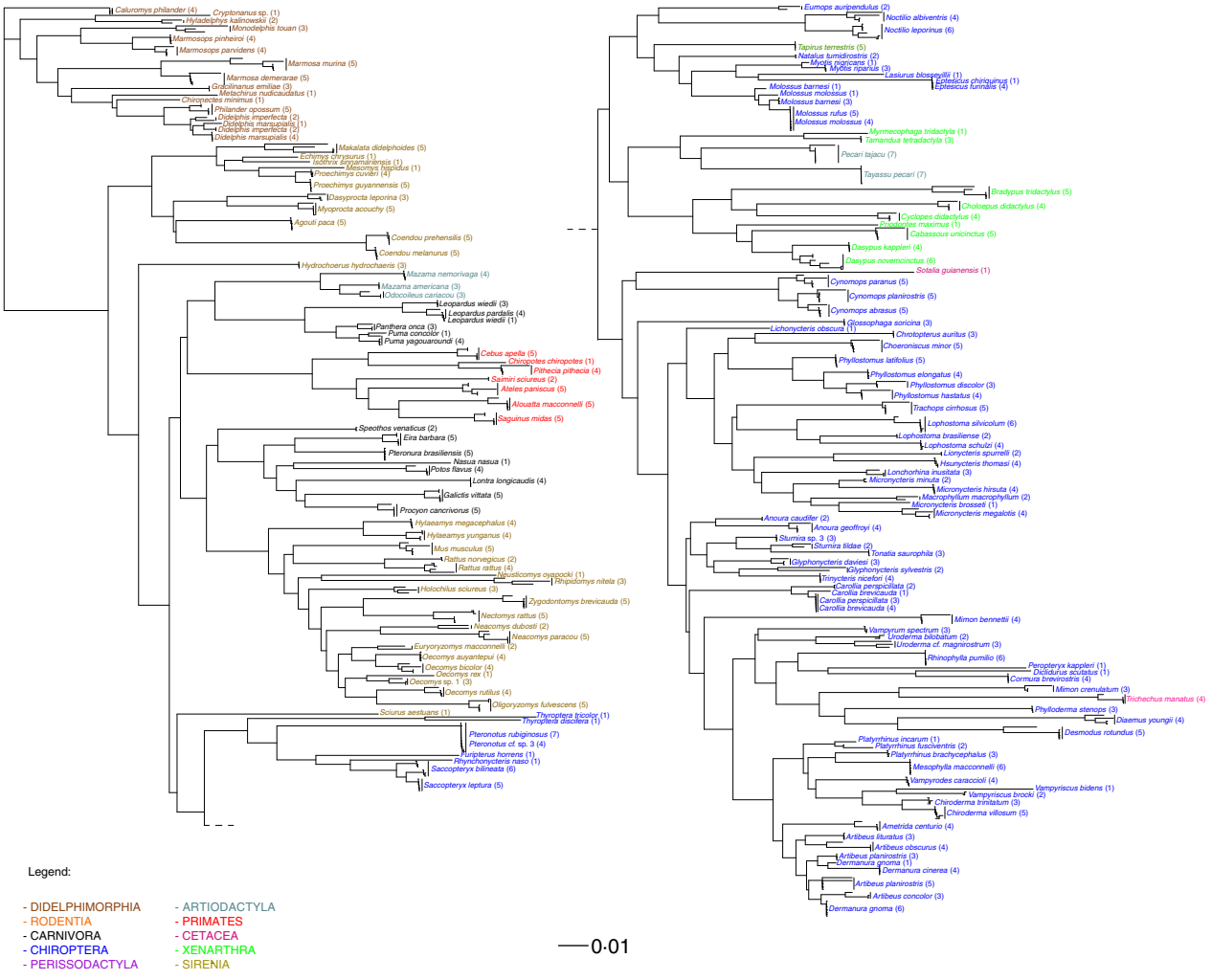

Fig. 4. Neighbour-joining tree based on raw distances computed from pairwise alignments of the 12S-V5 metabarcode. Numbers in brackets indicate the number of specimens sequenced per species. Species names are coloured by order.

© 2017 The Authors. Methods in Ecology and Evolution © 2017 British Ecological Society,

Table 1. Results of taxonomic assignations using ECOTAG: each specimen was considered as unknown while all other sequences were used as references. Percentages of identifications that were made at each taxonomic rank are indicated, as well as the percentage of specimens that could not be identified because they did not find a close match in the reference library

		Proportion of the identifications made at given taxonomic rank					
Metabarcode	Reference library	Species	Genus	Family	Order	Not identified	Overall error rate‡
12S-V5	Comprehensive*	90·0	9·5	0·0	0·0	0·5	0·4
	No conspecific†	23·6	10·5	0·7	0·0	65·2	26·9
COI_minibarcode	Comprehensive*	87·7	4·7	0·0	0·0	7·6	1·1
	No conspecific†	10·1	0·0	0·0	0·0	89·9	10·1

*All specimens have at least one conspecific to match against in the reference library.
†To assess the effect of potential gaps in the database, we removed every conspecific from the references before taxonomic assignations.
‡Percentage of specimens that were wrongly identified.

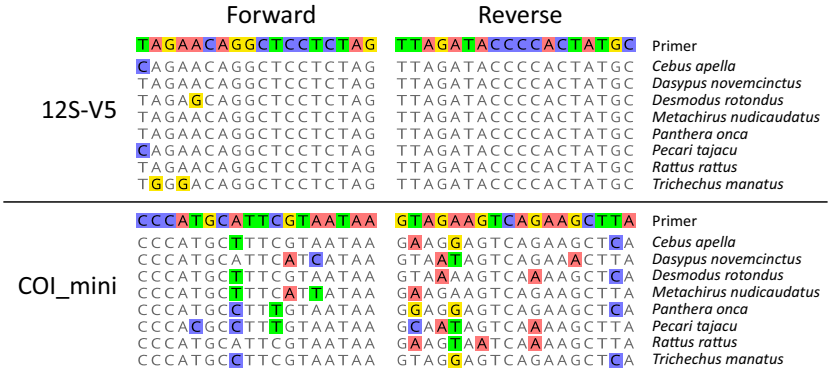

Fig. 5. Alignment of 12S-V5 and COI_minimam primers with reference sequences of mammals belonging to distinct orders. Primer mismatches are highlighted in the reference sequences.

were not identified because they did not find a close match, as it was expected. However, 23·6% of the sequences were falsely assigned at the specific level. 10·5% of the sequences were assigned at the generic rank with an error rate of 41%, corresponding to confusion between the genera *Artibeus* and *Dermanura*, *Didelphis* and *Philander* and *Puma* and *Panthera*. Finally, four sequences were assigned at the family level, all correctly. The overall proportion of the specimens that were assigned to a wrong taxon was 26·9%. If identifications made at the specific level were only considered at the generic level, they would have been correct in 93·9% of cases, and the overall proportion of the misidentified specimens would have been lowered to 4·7%.

Thirty blood-fed specimens, including four mosquitoes and 26 sand flies, were collected in French Guiana. Amplification and sequencing of the 12S-V5 marker was successful for 27 of the specimens, and allowed the identification of eight mammal species belonging to five distinct orders: Primate, Didelphimorphia, Rodentia, Carnivora and Xenarthra (see Supporting Information for details).

COMPARISON WITH COI

A total of 569 COI sequences representing 138 mammal species found in French Guiana were retrieved from BOLD. The newly designed PCR primers amplified a 102-pb-long fragment, and had similar theoretical melting temperatures (COI_minimam_F: 5'-CCCATGCATTCGTAATAA-3'; COI_minimam_R: 5'-GTAGAAGTCAGAAGCTTA-3'). The number of mismatches per reference sequence ranged from 0 to 5 for both primers for a mean of 2·13 and 2·2 for primer F and R, respectively. No reference sequence had zero mismatch with both primers. To visualize mismatching patterns of 12S and COI primers, an alignment of each primer pair with reference sequences of mammals from various orders is shown in Fig. 5. Assessment of the COI_minimam taxonomic resolution revealed that with a comprehensive reference database, 87·7 and 4·7% of sequences would have been assigned at the specific and generic level, respectively while, 7·6% would have remained unidentified, for an overall error rate of 1·1% (Table 1). With an incomplete reference database, 10·1% of the sequences would have been wrongly identified at the specific level, while the all others would have remained unidentified (i.e. an overall error rate of 10·1%).

Discussion

By enabling the identification of species from degraded DNA contained in environmental samples, DNA metabarcoding has opened great avenues for the study of vertebrates' communities. It has already been proved successful in various applications such as the characterization of the present or ancient terrestrial fauna from the soil (Andersen *et al.* 2012; Giguet-Covex *et al.* 2014), aquatic communities from water (Valentini *et al.* 2016) or feeding behaviours from faeces (De Barba *et al.* 2014). Because the standard COI barcode is not adapted for metabarcoding (Deagle *et al.* 2014), these studies relied on the development of other markers.

© 2017 The Authors. Methods in Ecology and Evolution © 2017 British Ecological Society,

First, we show that the 12S-V5 primer binding sites are extremely conserved among the mammal species included in our database. Second, we show that the marker allows reliable and precise identifications of mammals, which was further highlighted by the successful analysis of dipteran blood meals. Using a similarity cut-off of 97%, almost all specimens could be correctly identified and more than 90% of these identifications were made at the specific level, while the others were made at the generic level. Only two sequences were assigned to a wrong taxon (false-positive errors), while three specimens were not identified because their closest match did not reach the similarity cut-off (false-negative errors). This level of accuracy is remarkable regarding that the marker is only 100 bp long and can be amplified with the same PCR primers virtually in all vertebrates (Riaz *et al*. 2011). Our results emphasize that the quality of taxonomic assignations is largely dependent on the comprehensiveness of the reference database. Indeed, when all conspecific sequences were removed prior to taxonomic assignations, almost 30% of the specimens were assigned to a wrong taxon. This error rate is largely dependent on the choice of the similarity cut-off. A more stringent (higher) cut-off would have resulted in a lower false-positive error rate, but also in a higher false-negative error rate. The wide overlap observed between intraspecific and interspecific genetic distances, and thus, the absence of a clear barcoding gap precludes the existence of a perfect similarity cut-off. The choice of an optimum was not in the scope of this work. However, we have shown that bringing the taxonomic assignations to the generic rank restores the reliability of the identifications when the species of the query is not represented in the reference library. Therefore, besides using a stringent similarity threshold, we recommend to consider taxonomic assignations only at the generic level when using the 12S-V5 metabarcode with an incomplete reference library. The importance of the comprehensiveness of reference libraries to avoid erroneous identifications and the difficulty to define similarity cut-off because of the overlapping intra and interspecific genetic distances has already been highlighted for classical COI barcoding (Meyer *et al*. 2008; Puillandre *et al*. 2009). This should be even more significant for shorter and less discriminant metabarcodes. Numerous studies provide thoroughly sampled COI reference libraries together with an evaluation of barcoding in specific animal groups. On the contrary, DNA metabarcoding studies usually rely on public databases, and the current literature mainly focuses on biomolecular and bioinformatic issues, such as the management of PCR and sequencing artefacts, rather than providing the evaluation of metabarcodes accuracy based on comprehensive reference libraries. This is understandable because metabarcodes frequently target very wide taxonomic ranges or taxa in which a large proportion of species are unknown. In addition, there exists no real standard for metabarcoding markers (at the exception of bacteria and fungi), which impedes the creation of a single collaborative and well-curated reference database as the BOLD for barcoding (Pompanon & Samadi 2015). Nevertheless, we argue that these limitations should be considered seriously when interpreting metabarcoding results.

Finally, we compared the theoretical performances of the 12S-V5 marker with that of a newly designed COI metabarcode. The COI reference database retrieved on BOLD contained similar number of sequences and taxonomic coverage than the 12S reference database constituted in this study, which allowed relevant comparisons. The COI_minimam was selected using the same procedure than for the 12S-V5 marker. However, it is noteworthy that 12S-V5 primers were designed to amplify DNA of all vertebrates, whereas the COI_minimam was specifically directed to Amazonian mammals, which represent a rather conservative approach for our comparison. We show that while both metabarcodes provide comparable taxonomic resolution, the COI_minimam primers present high rates of primer binding site mismatches (Fig. 5), constituting a serious disadvantage for metabarcoding studies.

Besides providing reliable and comprehensive molecular data for the identification of mammals in French Guiana and more largely for the northern Amazon region, our study emphasizes that great caution should be taken concerning metabarcoding results based on scarce reference libraries and that molecular identifications should be trusted only using a stringent similarity threshold and at appropriate taxonomic ranks. Our results also provide novel evidence in support for the use of ribosomal markers in metabarcoding studies.

Authors' contributions

J.M., AL.B. and A.K. designed the study. B.d.T. and F.C. provided the samples. A.K., M.H., B.d.T. and S.V. performed the laboratory work. A.K. analysed the data. A.K. prepared the manuscript and all authors contributed in its improvement.

Acknowledgements

This work was supported by 'Investissement d'Avenir' grants managed by Agence Nationale de la Recherche (CEBA: ANR-10-LABX-25-01; TULIP: ANR-10-LABX-41, ANR-11-IDEX-0002-02) as well as project METABAR (ANR-11-BSV7-0020). Many samples used in this study are part of the JAGUARS collection, supported by DEAL Guyane, Collectivité Territoriale de Guyane, and European ERDF funds. Most of COI barcoding was done within the Guyamazon II project project 'Biodiversidade e zoogeografia de pequenos mamíferos no escudo das Guianas' which was funded by Ambassade de France in Brazil, IRD/AIRD (France), the Collectivité Territoriale de Guyane, CIRAD, FAPEAM, FAPEMA, and FAPEAP. We would like to thank Frederic Delsuc and Fidel Botero-Castro (Institut des Sciences de l'Evolution de Montpellier) for providing sequences of Xenarthra and Chiroptera, as well as Pierre Taberlet and Eric Coissac for fruitful discussions about metabarcoding.

Data accessibility

All sequences were deposited in GenBank (accessions: KX381203–KX381784).

References

Andersen, K., Bird, K.L., Rasmussen, M., Haile, J., Breuning-Madsen, H., Kjær, K.H., Orlando, L., Gilbert, M.T.P. & Willerslev, E. (2012) Meta-barcoding of 'dirt' DNA from soil reflects vertebrate biodiversity. *Molecular Ecology*, **21**, 1966–1979.

Binladen, J., Gilbert, M.T.P., Bollback, J.P., Panitz, F., Bendixen, C., Nielsen, R. & Willerslev, E. (2007) The use of coded PCR primers enables high-throughput sequencing of multiple homolog amplification products by 454 parallel sequencing. *PLoS ONE*, **2**, e197.

© 2017 The Authors. Methods in Ecology and Evolution © 2017 British Ecological Society,

Borisenko, A.V., Lim, B.K., Ivanova, N.V., Hanner, R.H. & Hebert, P.D.N. (2008) DNA barcoding in surveys of small mammal communities: a field study in Suriname. *Molecular Ecology Resources*, **8**, 471–479.

Boyer, F., Mercier, C., Bonin, A., Le Bras, Y., Taberlet, P. & Coissac, E. (2016) obitools: a unix-inspired software package for DNA metabarcoding. *Molecular Ecology Resources*, **16**, 176–182.

Brown, S.D., Collins, R.A., Boyer, S., Lefort, M.-C., Malumbres-Olarte, J., Vink, C.J. & Cruickshank, R.H. (2012) Spider: an R package for the analysis of species identity and evolution, with particular reference to DNA barcoding. *Molecular Ecology Resources*, **12**, 562–565.

Bru, D., Martin-Laurent, F. & Philippot, L. (2008) Quantification of the detrimental effect of a single primer-template mismatch by real-time PCR using the 16S rRNA gene as an example. *Applied and Environmental Microbiology*, **74**, 1660–1663.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K. & Madden, T.L. (2009) BLAST+: architecture and applications. *BMC Bioinformatics*, **10**, 421.

Catzeflis, F. (2015). Liste des Mammifères de Guyane. *Nature guyanaise: 50 ans de progrès et de souvenirs* (ed. L. Sanite), pp. 226–239. Editions Orphie, Saint-Denis de la Réunion.

Clarke, L.J., Soubrier, J., Weyrich, L.S. & Cooper, A. (2014) Environmental metabarcodes for insects: *in silico* PCR reveals potential for taxonomic bias. *Molecular Ecology Resources*, **14**, 1160–1170.

Coghlan, M.L., White, N.E., Murray, D.C., Houston, J., Rutherford, W., Bellgard, M.I., Haile, J. & Bunce, M. (2013) Metabarcoding avian diets at airports: implications for birdstrike hazard management planning. *Investigative Genetics*, **4**, 27.

De Barba, M., Adams, J.R., Goldberg, C.S., Stansbury, C.R., Arias, D., Cisneros, R. & Waits, L.P. (2014) Molecular species identification for multiple carnivores. *Conservation Genetics Resources*, **6**, 821–824.

Deagle, B.E., Jarman, S.N., Coissac, E., Pompanon, F. & Taberlet, P. (2014) DNA metabarcoding and the cytochrome c oxidase subunit I marker: not a perfect match. *Biology Letters*, **10**, 20140562.

Ficetola, G.F., Miaud, C., Pompanon, F. & Taberlet, P. (2008) Species detection using environmental DNA from water samples. *Biology Letters*, **4**, 423–425.

Gibb, G.C., Condamine, F.L., Kuch, M., Enk, J., Moraes-Barros, N., Superina, M., Poinar, H.N. & Delsuc, F. (2016) Shotgun mitogenomics provides a reference phylogenetic framework and timescale for living Xenarthrans. *Molecular Biology and Evolution*, **33**, 621–642.

Giguet-Covex, C., Pansu, J., Arnaud, F. *et al.* (2014) Long livestock farming history and human landscape shaping revealed by lake sediment DNA. *Nature Communications*, **5**, 3211.

Hebert, P.D., Cywinska, A., Ball, S.L. *et al.* (2003) Biological identifications through DNA barcodes. *Proceedings of the Royal Society of London B: Biological Sciences*, **270**, 313–321.

Huson, D.H., Auch, A.F., Qi, J. & Schuster, S.C. (2007) MEGAN analysis of metagenomic data. *Genome Research*, **17**, 377–386.

Kartzinel, T.R., Chen, P.A., Coverdale, T.C., Erickson, D.L., Kress, W.J., Kuzmina, M.L., Rubenstein, D.I., Wang, W. & Pringle, R.M. (2015) DNA metabarcoding illuminates dietary niche partitioning by African large herbivores. *Proceedings of the National Academy of Sciences*, **112**, 8019–8024.

Kimura, M. (1980) A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of Molecular Evolution*, **16**, 111–120.

Kocher, A., Gantier, J.-C., Gaborit, P. *et al.* (2016) Vector soup: high-throughput identification of Neotropical phlebotomine sand flies using metabarcoding. *Molecular Ecology Resources*, doi: 10.1111/1755-0998.12556. (in press)

Lim, B.K. (2012) Biogeography of mammals from the Guianas of South America. *Bones, Clones and Biomes: The History and Geography of Recent Neotropical Mammals* (eds B.D. Patterson & L.P. Costa), pp. 230–258. University of Chicago Press, Chicago, IL, USA.

Mariette, J., Escudié, F., Allias, N., Salin, G., Noirot, C., Thomas, S. & Klopp, C. (2012) NG6: integrated next generation sequencing storage and processing environment. *BMC Genomics*, **13**, 462.

Meyer, C.P. & Paulay, G. (2005) DNA barcoding: error rates based on comprehensive sampling. *PLoS Biology*, **3**, e422.

Meyer, F., Paarmann, D., D'Souza, M. *et al.* (2008) The metagenomics RAST server – a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics*, **9**, 386.

Paradis, E., Claude, J. & Strimmer, K. (2004) APE: analyses of phylogenetics and evolution in R language. *Bioinformatics*, **20**, 289–290.

Pompanon, F. & Samadi, S. (2015) Next generation sequencing for characterizing biodiversity: promises and challenges. *Genetica*, **143**, 133–138.

Puillandre, N., Strong, E.E., Bouchet, P., Boisselier, M.-C., Couloux, A. & Samadi, S. (2009) Identifying gastropod spawn from DNA barcodes: possible but not yet practicable. *Molecular Ecology Resources*, **9**, 1311–1321.

R Core Team (2014) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Riaz, T., Shehzad, W., Viari, A., Pompanon, F., Taberlet, P. & Coissac, E. (2011) ecoPrimers: inference of new DNA barcode markers from whole genome sequence analysis. *Nucleic Acids Research*, **39**, 1–11.

Saitou, N. & Nei, M. (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, **4**, 406–425.

Srivathsan, A. & Meier, R. (2012) On the inappropriate use of Kimura-2-parameter (K2P) divergences in the DNA-barcoding literature. *Cladistics*, **28**, 190–194.

Taberlet, P., Coissac, E., Pompanon, F., Brochmann, C. & Willerslev, E. (2012) Towards next-generation biodiversity assessment using DNA metabarcoding. *Molecular Ecology*, **21**, 2045–2050.

Valentini, A., Taberlet, P., Miaud, C. *et al.* (2016) Next-generation monitoring of aquatic biodiversity using environmental DNA metabarcoding. *Molecular Ecology*, **25**, 929–942.

Yu, D.W., Ji, Y., Emerson, B.C., Wang, X., Ye, C., Yang, C. & Ding, Z. (2012) Biodiversity soup: metabarcoding of arthropods for rapid biodiversity assessment and biomonitoring. *Methods in Ecology and Evolution*, **3**, 613–623.

Received 2 October 2016; accepted 16 December 2016
Handling Editor: Oscar Gaggiotti

Supporting Information

Details of electronic Supporting Information are provided below.

Data S1. List of the specimens used in this study and along with detail information.

Data S2. Bash script used for the analysis of sequencing data with the OBITOOLS.

Data S3. Dipteran blood meals analyses procedure and results.

© 2017 The Authors. Methods in Ecology and Evolution © 2017 British Ecological Society, *Methods in Ecology and Evolution*