An *In silico* approach for the evaluation of DNA barcodes

Gentile Francesco Ficetola¹,²,³*, Eric Coissac¹*, Stéphanie Zundel¹, Tiayyba Riaz¹, Wasim Shehzad¹, Julien Bessière¹, Pierre Taberlet¹, François Pompanon¹

Abstract

**Background:** DNA barcoding is a key tool for assessing biodiversity in both taxonomic and environmental studies. Essential features of barcodes include their applicability to a wide spectrum of taxa and their ability to identify even closely related species. Several DNA regions have been proposed as barcodes and the region selected strongly influences the output of a study. However, formal comparisons between barcodes remained limited until now. Here we present a standard method for evaluating barcode quality, based on the use of a new bioinformatic tool that performs *in silico* PCR over large databases. We illustrate this approach by comparing the taxonomic coverage and the resolution of several DNA regions already proposed for the barcoding of vertebrates. To assess the relationship between *in silico* and *in vitro* PCR, we also developed specific primers amplifying different species of Felidae, and we tested them using both kinds of PCR.

**Results:** Tests on specific primers confirmed the correspondence between *in silico* and *in vitro* PCR. Nevertheless, results of *in silico* and *in vitro* PCRs can be somehow different, also because tuning PCR conditions can increase the performance of primers with limited taxonomic coverage. The *in silico* evaluation of DNA barcodes showed a strong variation of taxonomic coverage (i.e., universality): barcodes based on highly degenerated primers and those corresponding to the conserved region of the Cyt-b showed the highest coverage. As expected, longer barcodes had a better resolution than shorter ones, which are however more convenient for ecological studies analysing environmental samples.

**Conclusions:** *In silico* PCR could be used to improve the performance of a study, by allowing the preliminary comparison of several DNA regions in order to identify the most appropriate barcode depending on the study aims.

**Background**

DNA barcoding, i.e., the identification of biological diversity using standardized DNA regions, has been demonstrated as a new, very useful approach to identify species [1]. Originally, DNA barcoding was proposed to assign an unambiguous tag to each species, giving to taxonomists a standard method for identification of specimens. In this context, it was also proposed that DNA barcoding is an opportunity to accelerate the discovery of new species [2-4]. Today, the fields of applications of this approach are broader. As example, DNA barcoding has been already used in biodiversity assessment, forensics, diet analysis and paleoecological studies [5-7].

In the former context, a portion of mitochondrial cytochrome c oxidase (*COI*) has been proposed as the standard barcode for animal identification [1,8]. Since then, other portions of DNA have been proposed as barcodes, because different DNA regions have different performances in some taxa (e.g., flowering plants [9,10]; amphibians [11]). If we consider the other applications of barcoding (*sensu lato* DNA barcoding, [6]), the necessity to limit the number of usable barcode loci for conserving the standard aspect of this method can be relaxed. In such a new context, multiple barcodes in different regions of the genome could be combined to improve identification, according to the taxon studied and to the aims of the research [9,10]. Therefore, the first step of a *sensu lato* barcoding study should be the selection of the best DNA region(s) to be used as barcode considering the aims of the study. The availability of large public sequence databases may allow comparing multiple potential barcodes and their properties before performing studies.

© 2010 Ficetola et al; licensee BioMed Central Ltd. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/2.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.
Among the properties of an ideal DNA barcode, high taxonomic coverage and high resolution are essential [6,12]. A high taxonomic coverage (also called universality) would allow the application of barcodes to a number of taxa as large as possible, including undescribed species. This constraint the DNA barcode region to have sufficiently conserved flanking regions enabling the design of universal primers. This is especially important for describing unknown biodiversity or diversity within environmental samples such as soils or faeces [6,7,13].

However, universality can be extremely difficult to achieve, because of the incomplete knowledge of genetic variation in poorly studied taxa [12]. The resolution capacity of a barcode is its ability to differentiate and identify species that relies on interspecific differences among DNA sequences [8,14]. Thus, the challenge for defining a barcode of good quality consists in finding a quite short and enough variable DNA sequence flanked by highly conserved regions. Depending of the application, the size, the taxonomic coverage or the resolution of the DNA barcode could be the most important characteristic to optimise [6].

This study proposes an explicit approach for comparing the performance of potential barcoding regions, which is based on 'in silico' PCRs performed over extensive databases, and on two indices that estimate the resolution capacity of the barcodes and the taxonomic coverage of the primers used for their amplification. As an example, we analysed several primers available from the literature that have been used in sensu lato barcoding studies [6] for the identification of Vertebrates species. First, we assessed the taxonomic coverage of several primer pairs by evaluating the proportion of species amplified in silico in a purposely designed database. Subsequently, we analyzed the GenBank sequences amplified by each primer pair, in order to evaluate the proportion of species correctly identified on the basis of their barcodes. We also used an in vitro analysis to validate the correspondence between in silico and real world PCR.

Methods
General strategy
First, we created a reference database representative of the mitochondrial genomes of all vertebrates, by retrieving from Genbank all the complete mitochondrial genomes of Vertebrates available (accession: September 2007). Subsequently, we randomly selected one sequence per species, to reduce the overrepresentation of a few species (e.g., humans, mouse, zebrafish etc.). We obtained a set of 814 mitochondrial genomes representative of the five major monophyletic clades of vertebrates [Chondrichthyes: 8 species; Actinopterigii: 385 species; Amphibia: 79 species; Sauropsida (= birds + "reptiles"): 133 species; Mammalia: 202 species; other taxa: 7 species]. Most of species were the unique representative of their genus and the database corresponded to 633 genera.

To analyze the performance of each primer pair studied, we first performed an in silico PCR on the reference database and we evaluated the taxonomic coverage of each primer pair as the proportion of amplified taxa. Then, we performed an in silico PCR on the whole GenBank, to evaluate the resolution of the amplified fragments that represents the proportion of unambiguously identified taxa. These properties were evaluated for the whole Vertebrates and for each of the five clades which compose it.

In Silico PCR
An in silico PCR consists in selecting in a database the sequences that match (i.e., exhibit similarity with) two PCR primers. The regions matching the two primers should be localised on the selected sequence in a way allowing PCR amplification, which forces the relative orientation of the matches and the distance between them. In order to simulate real PCR conditions, the in silico PCR algorithm should allow some mismatches between the primers and the target sequences. Standard sequence similarity assessment programs such as BLAST [15] are not suitable for such kind of analysis because the heuristic search they use is not efficient on short sequences. Moreover, a post processing of BLAST output should be performed to verify previously stated constraints. We have developed a program named ecoPCR that is based on the very efficient pattern matching algorithm Agrep [16]. This algorithm allows specifying the maximum count of mismatched positions between each primer and the target sequence, and to use the full IUPAC code (e.g., R for purines or Y for pyrimidines). It also allows specifying on which primer's specific positions mismatches are not tolerated, what is useful to force exact match on the 3' end of primers for simulating real PCR conditions. Moreover, to facilitate further analysis, ecoPCR output contains the taxonomic information for each sequence selected from the database. For the analyses presented in this article, we allowed two mismatches between each primer and the template, except on the last 3 bases of the 3' end of the primer. Analyses performed with 0, 1 or 3 mismatches led to similar conclusions (results not shown), even if the results were sometimes different (see discussion). This software was developed for Unix platforms and is freely available at http://www.grenoble.prabi.fr/trac/ecoPCR.

Measuring taxonomic coverage
To measure the taxonomic coverage of a primer pair, we defined a coverage index $B_{c}$ as the ratio between the number of amplified taxa for a specified taxonomic rank
(i.e., species for this analysis; genus or family can be specified as alternative taxonomic ranks) and the total number of taxa of the same level representing the studied clade in the reference sequence database. Bc can be computed from ecoPCR output file using the ecoTaxStat script.

Measuring resolution capacity

The resolution capacity of a barcode was estimated by an index measuring the ratio of unambiguously identified taxa for a given taxonomic level over the total number of tested taxa. A taxon unambiguously identified by a primer pair owns a barcode sequence associated to this pair that is not shared by any other taxa of the same taxonomic rank. To be computed, this definition can be formalized considering the mapping \( E, \) \( Img \) and \( E' \) between four concept sets: taxon \( (T) \), individual \( (I) \), barcode \( (B) \) and region \( (R) \) (for a full definition see figure 1). Considering the a taxon \( t \in T \) and a primer pair (barcode region) \( r \in R \) and using the mapping \( E, \) \( Img \) and \( E' \) we define the \( \Omega(t,r) \) set of all barcodes belonging to a taxon for a region:

\[
\Omega(t,r) = \text{Img}(E(t)) \cap E'(r)
\]

From the above description, we note the set of all individuals owning a barcode corresponding to a taxon as:

\[
\text{Img}^{-1}(\Omega) = \bigcup_i \text{Img}^{-1}(b_i, b_i \in \Omega)
\]

This allows defining an unambiguously identified taxon \( t \) by a barcode region \( r \) if and only if:

\[
\text{Img}^{-1}(\Omega(t,r)) = E(t)
\]

This defines a mapping \( \epsilon \) of \( T \) to \( R \) and allows to define the specificity index \( B_s \) as:

\[
B_s(r) = \left\{ \frac{|\{t/\epsilon(r)\}|}{|T|} \right\}
\]

\( B_s \) can be computed from an ecoPCR output file using the ecoTaxSpecificity script. ecoTaxSpecificity and ecoTaxStat scripts are parts of the OBITools python package freely available at http://www.grenoble.prabi.fr/trac/OBITools.

In a few cases, especially for Chondrichthyes, ecoPCR ran over the entire GenBank yielded only a small number of sequences. Thus, we calculated the resolution capacity of a barcode only when the primer pair amplified more than 10 species.

Correspondance between in vitro and in silico PCRs

Strict experimental validation of the electronic PCR realized over large databases would be extremely difficult, as it would require obtaining tissues from hundreds of species. Alternatively, specific primer pairs designed to amplify only one species can be used to confirm the correspondence between the results of ecoPCR and in vitro PCR. Therefore, we designed specific primers to amplify mitochondrial DNA of three species, using ecoPCR to test their specificity. Then, we cross-amplified the three species with each primer pairs with in vitro PCR to verify the ecoPCR predictions.

---

**Figure 1** Relationships between taxa, individuals, barcodes and regions as used in the \( B_s \) index estimation. In this example the taxon T1 is unambiguously identified by the R1 barcode region (green links) but the T2 is not well identified by the R1 region because this taxon share the B4 barcode region with the T3 taxon via the I6 individual (red links).
We considered three species of Asiatic Felidae: the Leopard (Panthera pardus); the Snow Leopard (Uncia uncia) and the Leopard cat (Prionailurus bengalensis). We designed specific primers for amplifying short sequences of mitochondrial 12S; this kind of primer pairs can be used to identify species from degraded DNA and remains, such as faeces. The three primer pairs were: (a) PantF, 5′-GTCACTAGATTTAGCT-3′, PantR, 5′-TGCCATATTTTATATTAAGTC-3′, designed to amplify the Leopard (amplified fragment: 120 bp); (b), UnciF, 5′-CTAAACCTAGATTTAGCT-3′, UnciR, 5′-CTCCTCTAGAGGGGTG-3′, designed to amplify the Snow Leopard (amplified fragment: 104 bp); (c) PrioF, 5′-CTCAACTTAGATTTAGCT-3′, PrioR, 5′-GGATGTAAAGCACCGCC-3′, designed to amplify the Cat Leopard (amplified fragment: 94 bp).

DNA was extracted from faeces using QiAamp DNA Stool Kit (Qiagen GmbH, Hilden, Germany). The PCRs were conducted in a 20 μl total volume with 8 mM Tris-HCl (pH 8.3), 40 mM KCl, 2 mM MgCl2, 0.2 μM of each primer, BSA (5 μg), 0.5 U of AmpliTaq Gold DNA polymerase (Applied Biosystems) and 2 ml of DNA extract. For all primers, the PCR programme included an initial 10 min denaturation step at 95°C, 45 cycles of denaturation at 95°C for 30 s and annealing at 53°C for 30 s. Samples of each of the three species were amplified with the three primer pairs, to verify in vitro the possibility of cross-amplification. We also tested cross-amplification ability of these primer pairs using ecoPCR, allowing two mismatches between each primer and the template, except on the last 3 bases of the 3′ end of the primer; subsequently, we simulated more relaxed PCR conditions [17] by allowing a larger number of mismatches.

Vertebrate primer pairs tested

The vertebrate primers tested (table 1) were selected in the bibliography as representative of the diversity of the strategies used for defining barcodes. Some of them (COI-1, COI-2, COI-3) were highly degenerated, in order to maximise the number of taxa amplified (i.e., the taxonomic coverage) [18]. Most of primers chosen amplified long sequences (> 500 bp) to maximize resolution, while some (e.g., Uni-Minibar, 16Smam) have been designed to amplify short sequences, to maximize the possibility of retrieving sequences from damaged/ancient DNA [19-21].

Results

Validation of in silico PCR

With in vitro PCR, each pair of specific primers amplified only the species for which it was designed: Pant
primers amplified Common Leopard only; Unci primers amplified Snow Leopard only, and Prio primers amplified Cat Leopard only (Figure 2). Cross-amplification through ecoPCR yielded identical results when allowing two mismatches. A more extensive analysis using ecoPCR, and allowing a larger number of mismatches (i.e., simulating more relaxed PCR conditions), shows that Pant primers require at least 3 mismatches for cross-amplifying *Uncia uncia*. Similarly, Unci and Prio primers require at least 4 mismatches for cross-amplifying other species.

**Evaluation of vertebrate primer pairs: Taxonomic coverage**

The primer pairs tested showed very different taxonomic coverage. Overall, COI-2, 16Sr and 16Sr2 were the primers with the highest percentages of species amplified (95, 90 and 93% of vertebrates amplified, respectively; Figure 3, table 2). Following our *in silico* PCRs, the primers with the lowest coverage corresponded to Uni-Minibar, COI-1, COI-2H, MCB and cytM. The primers also differed in their performance in amplifying the major clades of vertebrates. For example, COI-3 had the highest amplification rate in Chondrichthyes, while it amplified only 32% of the mammals. Conversely, 16Smam amplified most of the mammals, but failed in the amplification of Chondrichthyes (Figure 3, table 2). Nevertheless, in a similar way to how modifying the annealing temperature influences *in vitro* PCR [17], the number of electronically amplified species can be quickly increased by allowing a larger number of mismatches (Figure 4). For example, with primers Uni-Minibar, the proportion of amplified species reached 98% with eight tolerated mismatches (Figure 4).

**Resolution capacity of barcode regions**

When tested over the entire GenBank, most of the primer pairs had a very high resolution capacity, indicated by a high B_s index (Figure 5; table 2). We did not calculate B_s for primers Uni-Minibar and COI-2H because of the low number of species amplified with the settings used for this analysis (see discussion). Only the 16Smam primer pair, which amplifies a very short sequence (140 bp), had B_s < 85%. B_s was ≥ 90% for all other primer pairs and even > 97% for 16Sr and 16Sr2 whatever the vertebrate clade analysed (Figure 3, table 2). Apart from a few cases (e.g., low resolution of cytM within Actinopterigii), the resolution capacity of all primer pairs was consistently high across all taxa tested. These B_s differences are not correlated with the number of Genbank sequences amplified (analysis over all vertebrates: Spearman’s correlation $r_S = -0.323$, $N = 8$, $p = 0.4$; the correlations between resolution and number of amplified sequences were not significant also within the monophyletic groups analysed).

The *in silico* PCRs performed over the entire GenBank always yielded sequences from the target mitochondrial region. None of the primers amplified sequences recorded as nuclear sequences in GenBank.

**Discussion**

The identification of universal primer pairs amplifying fragments with high resolution capacity is a major task.

---

**Figure 2** Capillary electrophoresis (QIAxcel System, Qiagen) showing the results of cross-amplification of three species of Felidae using three specific primers. A01: Unci primers, template DNA from *Uncia uncia*; A02: Unci primers, template DNA from *Panthera pardus*; A03: Unci primers, template DNA from *Prionailurus bengalensis*; A04: Pant primers, template DNA from *U. uncia*; A05: Pant primers, template DNA from *P. pardus*; A06: Pant primers, template DNA from *P. bengalensis*; A07: Prio primers, template DNA from *U. uncia*; A08: Prio primers, template DNA from *P. pardus*; A09: Prio primers, template DNA from *P. bengalensis*. The size in base pairs is indicated on the left and on the right.
of DNA barcoding, and can help the broad scale analysis of life on earth. However, some authors argued that it is impossible that a single short sequence will be enough to distinguish all members of all species [12]. In this context, explicit in silico approaches like the one presented in this study allow analysing the properties of different sets of primers, and identifying the most appropriate ones a priori.

In silico vs. real PCR

The real in vitro amplification pattern depends on PCR conditions. Controlling the PCR conditions can alter amplification results, and thus the taxonomic coverage of primers. For example, low annealing temperature and high concentration of MgCl₂ reduce the specificity of primers in real-world PCR, and can thus allow amplification of target sequences with a larger number of
mismatches in the primer regions [17]. Our \textit{in silico} analyses have been performed allowing two mismatches. These parameters correspond well to actual amplification at rather high annealing temperatures (Figure 2), in accordance with previously published environmental genetics studies [22]. Nevertheless, these stringent conditions probably lead us to predict more false negative results (non electronic amplification of amplifiable sequences) than false positive ones (electronic amplification of non amplifiable sequences). Increasing the authorized mismatches can simulate more relaxed conditions, but the strict relationship between electronic and experimental conditions cannot be formally described. On the other hand, stringent PCR conditions reduce the risk of amplifying unwanted regions of the genome (see below), particularly when using degenerate primers. Furthermore, our study focused on \textit{sensu lato} barcode primer pairs. These studies often amplify DNA extracted from environmental samples, which may represent a mix of the DNA of several taxa [6]. Considering this, primers and PCR conditions must be as specific as possible, because the rare species with a low number of mismatches in the primer region (Figure 4) are expected to be overamplified and overrepresented in the PCR products, while species that are present, but with a higher number of mismatches, may not be amplified enough to yield sequences. Therefore, "ideal" primers would have a constantly low number of mismatches, leading to a less biased estimate of species presence.

EcoPCR can also be used to simulate less stringent PCR conditions, allowing more mismatches. With this approach, primers can amplify a much larger number of species (Figure 3). For example, in our stringent \textit{in silico} analysis the primers Uni-Minibar showed limited taxonomic coverage, and amplified very few vertebrates (table 2). Conversely, the PCRs performed by Meusnier \textit{et al.} [20] showed that these primers can amplify nearly 100% of fish and Amphibians, at an annealing temperature of 46°C. Results coherent with Meusnier \textit{et al.} [20] can be obtained using ecoPCR by allowing a large number of mismatches (up to eight) (Figure 4). Taking into account all these considerations, we have to assume that the taxonomic coverage \( B_c \) estimated from ecoPCR is not an exact value, but it reflects the relative capacity of primer pairs to amplify a broad variety of taxa. For example, the fact that 16Sr amplifies a much larger number of species of amphibians than COI-2H [[11,23], see also [24] for a different approach] was correctly predicted by \textit{in silico} analyses (see Figure 2, table 2).

Pseudogenes are a further potential issue in barcoding analysis; our approach may be affected by this trap. For instance, in our analyses none of the primers amplified nuclear sequences. However, nuclear sequences are underrepresented in Genbank; furthermore, the \textit{in silico} amplification of pseudogenes would require the presence of a target nuclear sequence and both the corresponding primer regions, i.e., a good coverage of nuclear genome. Therefore it is difficult that ecoPCR hits nuclear pseudogenes, which can nevertheless be amplified by \textit{in vitro} PCR, particularly under relaxed (e.g., low annealing temperature) conditions. Another potential issue of our approach is that the adjoining primer regions of sequences submitted to the databases are not a queryable portion of the database, therefore limiting the number of sequences obtained when ecoPCR is run over the

\[ \text{Figure 4 Mismatches between Uni-Minibar primers and vertebrate sequences. The histograms show the distribution of mismatch counts between (A) direct Uni-Minibar primer or (B) reverse Uni-Minibar primer and their target loci on mitochondrial DNA, as revealed by a ecoPCR run using Uni-Minibar primers to amplify our mitochondrial reference database. For this run, 8 mismatches were tolerated for each primer. The red curve and the associated right axis represent the cumulative fraction of amplified sequences with less that } m \text{ mismatch. We can observe than a few species present a small count of mismatches; these sequences with a few mismatches are expected to be advantageously amplified in a DNA mix containing multiple species.} \]
entire GenBank. To partially address this issue, the assessment of taxonomic coverage was performed on species for which the whole mitochondrial genome was available, and therefore both target sequences and flanking regions are present. The increasing availability of whole mitochondrial genomes due to the improvement sequencing technologies, and the rising of phylogenomics may reduce this limitation in the next future.

The correspondence between in silico and real PCR is certainly more accurate for the resolution capacity, still potential sources of bias remain. Our approach is based on the analysis of all the sequences deposited in GenBank, i.e., including thousands of vertebrate species in the example developed here. Assuming that all GenBank sequences are assigned to the correct species in the database, such approach uses the same kind of information than large scale barcoding studies. Clearly, the availability of sequences in different clades depends on the previous use of markers. For example, GenBank includes a very large number of COI sequences for Actinopterigii, while most of the mitochondrial sequences of mammals and amphibians are 16S. Furthermore, annotation errors are present in Genbank [25], and the error rate might be clade dependent. The $B_5$ index is sensible to these errors, leading to an underestimation of $B_5$; therefore, as for $B_C$ previously, $B_5$ should be considered as a relative measure of primer performance.

Comparison of vertebrate barcodes

Universality is a key feature of barcodes, and several strategies exist that can increase the taxonomic coverage of primer pairs. One strategy consists in making cocktails of degenerate primers. For example, the COI-2 primer pair [18] had one of the highest taxonomic coverages (figure 2). A predictable drawback of degenerate primers is a limited specificity with regards to the target DNA sequence amplified. However, our in silico PCRs performed on the whole GenBank did not amplify incorrect regions. All sequences amplified by the COI-2 primer pair were labelled in GenBank as mitochondrial COI, suggesting that these primers maintained enough specificity.

An alternative strategy consists in designing universal primers on highly conserved regions. This strategy has been used for example on the 16S, that exhibits some highly conserved regions in vertebrates [26]. The primers amplifying the 16S [[27,28]; this study] were very powerful, and had the highest taxonomic coverage and resolution capacity in vertebrates (Figure 2, Figure 3, table 2). The 16S region has been investigated as an alternate barcode locus for amphibians [11] but COI has not been rejected [24]. Some studies advocated that 16S has a too low rate of molecular evolution, and thus does not hold enough interspecific variation for a correct species identification [1]. Our analysis suggests that, at least in vertebrates, 16S has the same resolution capacity as COI, when using sequences with comparable length (500-600 bp), and therefore can be a good candidate site for barcoding. Nevertheless, the good performance of 16S observed in vertebrates may not be valid in other taxa; our in silico approach can be a key tool to analyse this possibility.

Long barcodes (500-600 bp) like the standard COI and 16S barcodes have a high resolution capacity, and are ideal candidates, for example, to unambiguously identify taxa in the context of the original DNA barcoding
usage. However, studies analysing environmental samples or degraded DNA require the use of shorter DNA fragments [6,7,13,20,22,29] even though those smaller regions include less information. We have included in our analysis two primer pairs amplifying short sequences that can be used for such analyses: Uni-Minibar [20] and 16S mam [21], which amplify sequences of 130-140 bp. Our analysis did not amplify enough sequences to evaluate the overall performance of Uni-Minibar, but allowed estimating the taxonomic coverage of 16S mam, which was very high for mammals (i.e., the taxon for which the primers have been designed), and lower for the other clades (Figure 2). This short barcode had the lowest resolution capacity for identification at the species level (Figure 3). However, in many cases species identification is not needed in ecological barcoding, as information on the genus or family can be already valuable [6,7,13,29]. Indeed, the resolution of 16S mam was much higher if the aim was the identification at the genus or family level (resolution capacity of 96% and 100%, respectively; results not shown).

Our analysis focused on vertebrates, because several primers have been proposed for their sensu lato barcoding. Furthermore, the in silico assessment of primers strongly depends on the sequences in online databases; vertebrates are the phylum best covered by available sequences, therefore they are the ideal focus of a methodological analysis. Nevertheless, biodiversity on Earth is dominated by other phyla, such as arthropods and molluscs: The evaluation method described here can be applied to these taxa and to any other ones, considering that the precision of the estimated B_S and B_C indices is directly linked to the amount and the quality of available sequences in public database corresponding to the studied clade.

Conclusion

Based on our in silico analyses, the different barcodes tested showed dissimilar adequacy to be used according to the five clades of vertebrates studied. If we consider all possible applications of sensu lato barcoding, no single barcode could be identified as the best for all vertebrates. The primers amplifying COI-2 showed the highest taxonomic coverage in Actinopterygii and Saur-opsida, while those amplifying 16Sr/16Sr2 showed the highest coverage of Amphibians and Mammals (Figure 3, table 2). Furthermore, the barcodes with the highest taxonomic coverage and resolution capacity (i.e., COI-2, 16Sr, 16Sr2) amplified long fragments, which can make their application problematic for describing biodiversity within environmental samples. In such a context, it is useful to select a priori the barcode that best suited the research topic. Our in silico method can help identifying the most appropriate barcode according to different aims. Such formal approach, which is possible thanks to the availability of bioinformatics tools and large public databases, can focus on target taxa or DNA regions and would make easier the validation of new barcodes by reducing the number of candidate primer pairs to be tested in vitro.

Acknowledgements

We thank three anonymous reviewers for constructive comments on an earlier version of the manuscript. GFF was funded by CNRS and LECA; this work has been partly supported by the Alocad project (ANR-06-PNRA-004-02) and by the European Project EcoChange (FP6-035866).

Author details

1Laboratoire d’Ecologie Alpine, CNRS UMR 5553, Université Joseph Fourier, BP 53, F-38041 Grenoble Cedex 9, France. 2Dipartimento di Biologia, Università degli Studi di Milano. Via Celona 26, 20133 Milano Italy. 3Dipartimento di Scienze dell’Ambiente e del Territorio, Università degli Studi di Milano Bicocca. Piazza della Scienza 1, 20126 Milano Italy.

Authors’ contributions

GFF, EC, PT and FP participated to the design of the study; JB and EC developed ecoPcr; TR and EC developed Bs and Bc indices, WS performed in vitro experiments; GFF and SZ performed the analyses; GFF, EC, PT and FP wrote the paper. All authors read and approved the final manuscript.

Received: 6 November 2009 Accepted: 16 July 2010
Published: 16 July 2010

References

1. Hebert PDN, Cywinska A, Ball SL, DeWaard JR. Biological identifications through DNA barcodes. Proc R Soc B 2003, 270:313-321.
2. Barber P, Boyce SL. Estimating diversity of Indo-Pacific coral reef stomatopods through DNA barcoding of stomatopod larvae. Proc R Soc B 2006, 273:2053-2061.
3. Hebert PDN, Penton EH, Bums JM, Janzen DH, Hallwachs W. Ten species in one: DNA barcoding reveals cryptic species in the neotropical skipper butterfly Astraptes fulgerator. Proc Natl Acad Sci USA 2004, 101:14812-14817.
4. Janzen DH, Hallwachs W, Blandin P, Bums JM, Cadiou JM, Chacon I, Dapkev T, Deans AR, Epstein ME, Espinoza B, et al. Integration of DNA barcoding into an ongoing inventory of complex tropical biodiversity. Mol Ecol Resour 2009, 9:1-26.
5. Willerslev E, Hansen AJ, Binladen J, Brand TB, Gilbert MTP, Shapiro B, Bunce M, Wulf C, Gilichinsky DA, Cooper A. Diverse plant and animal genetic records from Holocene and Pleistocene sediments. Science 2003, 300:791-795.
6. Valentini A, Pompfaffon F, Taberlet P. DNA barcoding for ecologists. Trends Ecol Evol 2009, 24:110-117.
7. Valentini A, Miquel C, Nawaz N, Bellemain E, Coissac E, Pompanon F, Geily L, Cruaud C, Nascetti G, Wilkinson P, et al. New perspectives in diet analysis based on DNA barcoding and parallel pyrosequencing: the inML approach. Mol Ecol Resour 2009, 9:51-60.
8. Hebert PDN, Ratnasingham S, DeWaard JR. Barcoding animal life: cytochrome c oxidase subunit 1 divergences among closely related species. Proc R Soc B 2003, 270:596-599.
9. Kress WJ, Wurdack KJ, Zimmer EA, Weigt LA, Janzen DH. Use of DNA barcodes to identify flowering plants. Proc Natl Acad Sci USA 2005, 102:8369-8374.
10. Hollingsworth PM, Forrest LL, Spouge JL, Hajibabaei M, Ratnasingham S, Bank van der M, Chase MW, Cowan RS, Erickson DL, Fazeekas AJ, et al. A DNA barcode for land plants. Proc Natl Acad Sci USA 2009, 106:12794-12797.
11. Vences M, Thomas M, Bonett RM, Vienes DR. Deciphering amphibian diversity through DNA barcoding: chances and challenges. Phil Trans R Soc B 2005, 360:1859-1868.
12. Rubinoff D, Cameron S, Will K. Are plant DNA barcodes a search for the Holy Grail? Trends Ecol Evol 2006, 21:1-2.
13. Deagle BE, Kirkwood R, Jarman SN. Analysis of Australian fur seal diet by pyrosequencing prey DNA in faeces. Mol Ecol 2009, 18:2022-2038.

14. Moritz C, Cicero C. DNA Barcoding: Promise and pitfalls. PLoS Biol 2004, 2:1529-1531.

15. Altschul SF, Madden TL, Schaffer AA, Zhang JH, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997, 25:3389-3402.

16. Wu S, Manber U. Agrep - a fast approximate pattern-matching tool. Proceedings of the Winter 1992 USENIX Conference San Francisco USA 20-24 Jan 1992 Berkeley, USA 1992, 153-162.

17. Ellich HA, Gelfand D, Drnisky JJ. Recent advances in the Polymerase Chain Reaction. Science 1991, 252:1643-1651.

18. Ivanova NV, Zemlak TS, Hanner RH, Hebert PDN. Universal primer cocktails for fish DNA barcoding. Mol Ecol Notes 2007, 7:544-548.

19. Hajibabaei M, Smith MA, Janzen DH, Rodrigues JJ, Whitfield JB, Hebert PDN. A minimalist barcode can identify a specimen whose DNA is degraded. Mol Ecol Notes 2006, 6:959-964.

20. Meunier I, Singer GAC, Landry J-F, Hickey DA, Hebert PDN, Hajibabaei M. A universal DNA mini-barcode for biodiversity analysis. BMC Genomics 2008, 9:214.

21. Taylor PG. Reproducibility of ancient DNA sequences from extinct pleistocene fauna. Mol Biol Evol 1996, 13:283-285.

22. Ficetola GF, Amaud C, Pompanon F, Taberlet P. Species detection using environmental DNA from water samples. Biol Lett 2008, 4:423-425.

23. Vences M, Thomas M, Meijden van der A, Chiari Y, Veites DR. Comparative performance of the 16S rRNA gene in DNA barcoding of amphibians. Front Zool 2005, 2:S.

24. Smith MA, Poyarkov NA, Hebert PDN. CO1 DNA barcoding amphibians: take the chance, meet the challenge. Mol Ecol Resour 2008, 8:235-246.

25. Schnoes AM, Brown SD, Dodeskii I, Sambit PC. Annotation error in public databases: Misannotation of molecular function in enzyme superfamilies. PLoS Comput Biol 2009, 5:

26. Kocher TD, Thomas WK, Meyer A, Edwards SV, Paabo S, Villablancan FX, Wilson AC. Dynamics of mitochondrial-DNA evolution in animals - Amplification and sequencing with conserved primers. Proc Natl Acad Sci USA 1989, 86:6196-6200.

27. Palumbi SR. Nucleic acids II: the polymerase chain reaction. Molecular Systematics Sunderland, Massachusetts: Sinauer & Associates Hills MD, Montz C, Mable BK 1996, 205-247.

28. Palumbi SR, Martin A, Romano S, McMullan WO, Stice L, Grabovski G. The simple fool's guide to PCR, ver 2 Honolulu: University of Hawaii 1991.

29. Pegard A, Miquel C, Valentini A, Coissac E, Bouvier F, Francois D, Taberlet P, Engel E, Pompanon F. Universal DNA-Based Methods for Assessing the Diet of Grazing Livestock and Wildlife from Feces. J Agric Food Chem 2009, 57:5700-5706.

30. Verma SK, Singh L. Novel universal primers establish identity of an enormous number of animal species for forensic application. Mol Ecol Notes 2003, 3:28-31.

31. Meyer R, Hoffelein C, Candraen U. Polymerase chain reaction restriction fragment length polymorphism analysis: A simple method for species identification in food. Journal of AOAC International 1995, 78:1542-1551.

doi:10.1186/1471-2164-11-434
Cite this article as: Ficetola et al.: An in silico approach for the evaluation of DNA barcodes. BMC Genomics 2010 11:434.

Submit your next manuscript to BioMed Central and take full advantage of:

- Convenient online submission
- Thorough peer review
- No space constraints or color figure charges
- Immediate publication on acceptance
- Inclusion in PubMed, CAS, Scopus and Google Scholar
- Research which is freely available for redistribution

Submit your manuscript at www.biomedcentral.com/submit