biology letters

rsbl.royalsocietypublishing.org

Opinion piece

Cite this article: Deagle BE, Jarman SN, Coissac E, Pompanon F, Taberlet P. 2014 DNA metabarcoding and the cytochrome *c* oxidase subunit I marker: not a perfect match. *Biol. Lett.* **10**: 20140562.
http://dx.doi.org/10.1098/rsbl.2014.0562

Received: 17 July 2014
Accepted: 19 August 2014

Subject Areas:
ecology, environmental science, molecular biology

Keywords:
DNA metabarcoding, DNA barcoding, cytochrome oxidase I

Author for correspondence:
Bruce E. Deagle
e-mail: bruce.deagle@aad.gov.au

Electronic supplementary material is available at http://dx.doi.org/10.1098/rsbl.2014.0562 or via http://rsbl.royalsocietypublishing.org.

Royal Society Publishing

Population genetics

DNA metabarcoding and the cytochrome *c* oxidase subunit I marker: not a perfect match

Bruce E. Deagle[1], Simon N. Jarman[1], Eric Coissac[2,3], François Pompanon[2,3] and Pierre Taberlet[2,3]

[1]203, Channel Highway, Kingston, Tasmania, Australia
[2]Université Grenoble Alpes, and [3]Centre National de la Recherche Scientifique (CNRS), Laboratoire d'Ecologie Alpine, F-38000 Grenoble, France

DNA metabarcoding enables efficient characterization of species composition in environmental DNA or bulk biodiversity samples, and this approach is making significant and unique contributions in the field of ecology. In metabarcoding of animals, the cytochrome *c* oxidase subunit I (COI) gene is frequently used as the marker of choice because no other genetic region can be found in taxonomically verified databases with sequences covering so many taxa. However, the accuracy of metabarcoding datasets is dependent on recovery of the targeted taxa using conserved amplification primers. We argue that COI does not contain suitably conserved regions for most amplicon-based metabarcoding applications. Marker selection deserves increased scrutiny and available marker choices should be broadened in order to maximize potential in this exciting field of research.

1. Introduction

Availability of affordable high-throughput DNA sequencing (HTS) has opened a new world of possibilities in DNA-based surveys of biodiversity. This approach is most advanced in the field of microbiology, where molecular taxonomy has a long tradition, and analyses now regularly use HTS to characterize markers for estimates of taxonomic as well as functional diversity. Amplified 'barcode' genes are also increasingly being used to identify plants, invertebrates and vertebrates present in DNA mixtures—obtained either by extracting total DNA from pooled specimens or from environmental samples (e.g. soil, water and faeces). This characterization of DNA barcodes from mixtures of DNA has been termed 'metabarcoding' [1,2].

Beyond the requirement for inexpensive and reliable sequence data, metabarcoding also needs a suitable marker. For standard DNA barcoding of single animal specimens, the Consortium for the Barcode of Life (CBOL) has adopted the mitochondrial cytochrome *c* oxidase subunit I (COI) gene. This marker has the required attributes: its variation usually allows species-level discrimination, it can be PCR amplified from most animals and the associated database now boasts millions of taxonomically verified DNA sequences. It seems like the obvious choice of marker in the nascent field of animal metabarcoding, and it has been used in many recent studies, including applications in biodiversity surveys, environmental monitoring and dietary studies (example studies provided in the electronic supplementary material).

2. So what is wrong with cytochrome *c* oxidase subunit I as a metabarcoding marker?

While COI can be amplified from an enormous range of species, it has always been acknowledged that primer binding sites within this protein-coding gene are not highly conserved. Mutations at many nucleotide positions do not change the

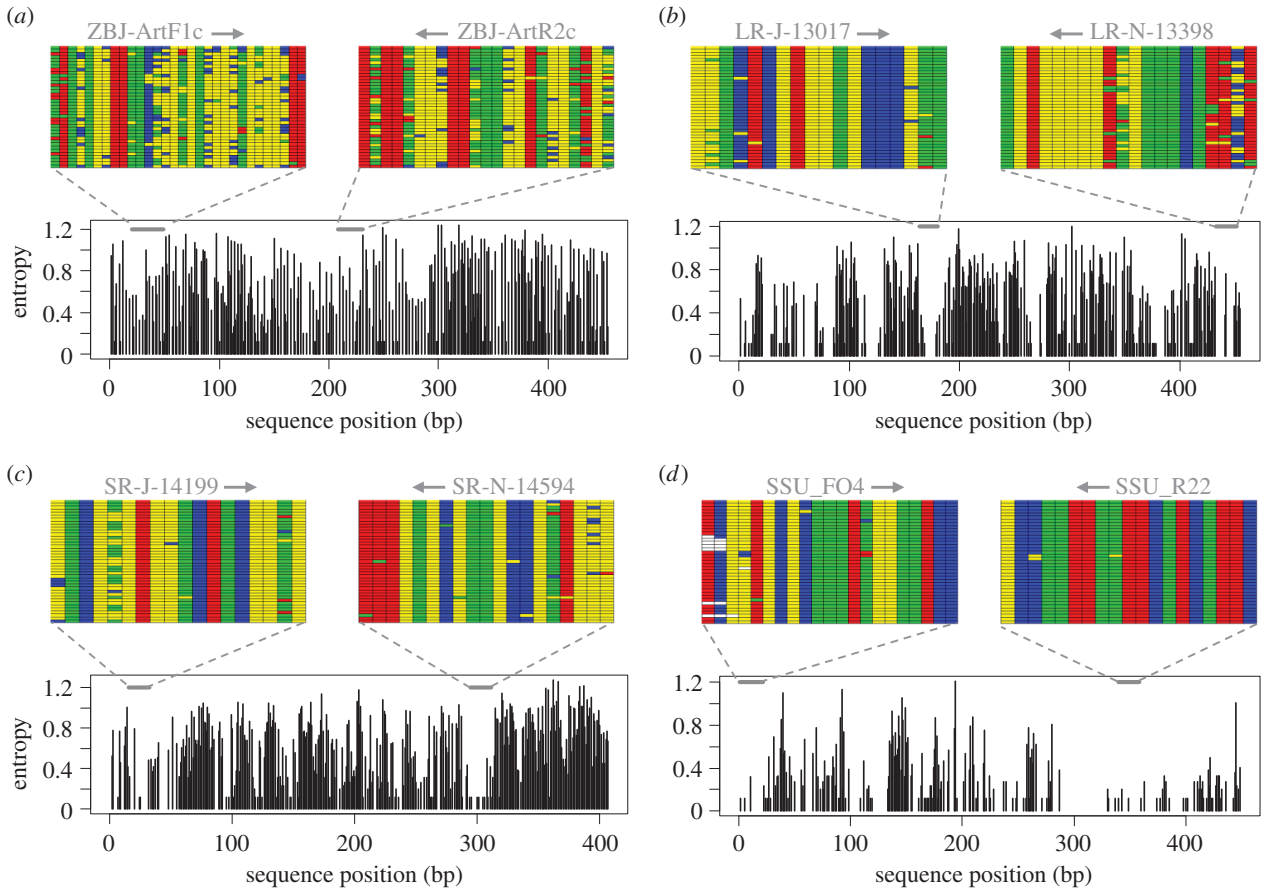

Figure 1. Variability of potential metabarcoding markers in representative insects (40 species from 25 different orders): (*a*) mtDNA COI (5′ region), (*b*) mtDNA 16S (5′ region), (*c*) mtDNA 12S and (*d*) nuclear 18S (5′ region). Data were extracted from full mtDNA datasets and comparable nuclear 18S rRNA gene sequences. Entropy represents a measure of variability at a given position and shading in highlighted primer sites shows the four nucleotides. The COI primers have been applied in over a dozen metabarcoding studies; see the electronic supplementary material for details. (Online version in colour.)

coded protein (usually the last base of the triplet code) and are less constrained by selection. Accordingly, a large number of primers have been designed for amplification of COI from various animal groups (currently more than 400 COI primers in the CBOL primer database). 'Universal' primers amplifying the COI barcode region have also been described, but *in silico* analysis shows they are poorly conserved ([3]; figure 1). Empirical studies indicate that this primer variability results in unreliable amplification when samples include species covering a broad taxonomic range (e.g. 44% success in more than 2000 initial amplifications; Moorea Biocode Project [4]). In standard DNA barcoding, it is possible to optimize protocols to get data from specimens that initially fail to amplify. However, when metabarcoding a DNA mixture, failed amplification of particular taxa is masked by the recovery of amplicons from other taxa present in the sample. This makes protocol optimization difficult. Furthermore, the recovery of some expected sequences gives false confidence in the resultant dataset.

Many microbial ecology studies have shown that although mismatched primers are able to amplify DNA from diverse bacterial genomes, targets without perfect homology amplify at lower and often unpredictable efficiency [5]. In some cases, even a single base mismatch can produce a 1000-fold underestimate of abundance [6], making some bacteria 'nearly undetectable' in HTS analysis of mock communities [7]. The use of cocktails with several primer variants can increase amplification success rates in standard DNA barcoding [4], but based on recent evaluations these are not a panacea for COI metabarcoding [2,8]. This is likely due to the fact that labile sites in COI primer binding regions diverge quickly (figure 2). Therefore, the number of primers required to account for variability, even between relatively closely related taxa, quickly becomes untenable. Furthermore, not all of these primer sequences will be effective at amplifying DNA (further discussion in the electronic supplementary material). A separate issue for COI metabarcode primer design is that variation at less constrained sites becomes saturated between distantly related taxa as a result of homoplasy (figure 2). This plateau in sequence divergence hinders development of group-specific primers (e.g. targeting all insects but excluding other terrestrial arthropods).

Notwithstanding these limitations, several COI primer sets have been developed specifically for metabarcoding. For example, a number of COI 'mini-barcoding' primers for amplifying short fragments recoverable from degraded template have been published even though primer sites vary among target species and alternative markers seem more suitable (figure 1). Metabarcoding primer cocktails have also been designed to amplify the full COI barcoding region in marine invertebrates, despite fewer than 50% of nucleotides at binding sites being conserved in the targeted taxa [4].

3. Is it best to accept biases and stick with standard barcode markers for metabarcoding?

It could be argued that biases introduced by differential COI primer binding are manageable if they are consistent across

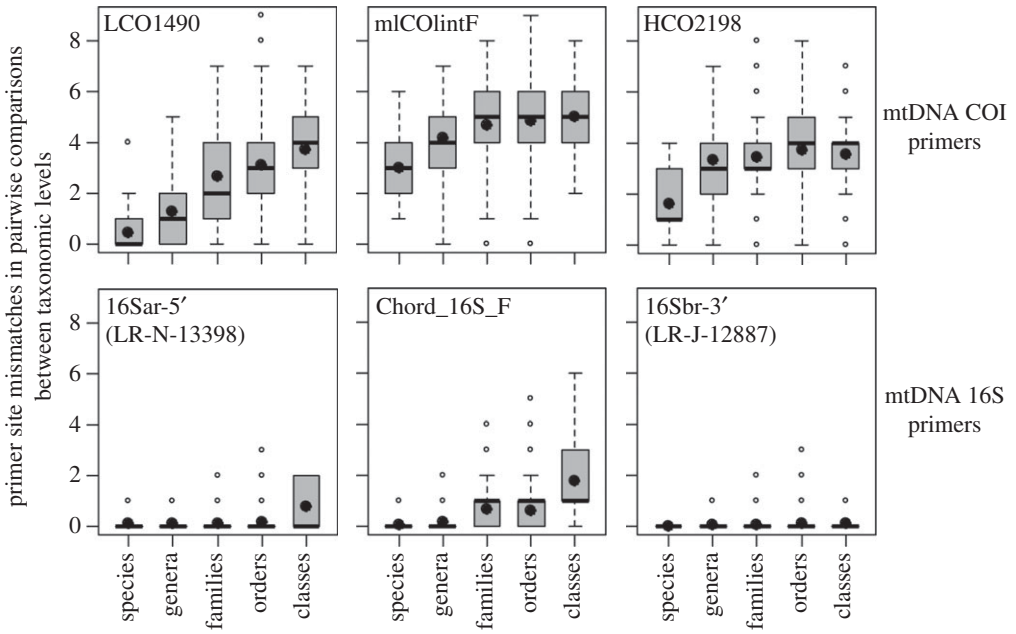

Figure 2. Variability in primer binding regions in two mtDNA markers (COI and 16S) at different taxonomic levels. Binding sites compared include those for primers commonly used to amplify a full-length COI barcode marker, those for an internal COI metabarcoding primer and analogous mtDNA 16S primers. Mismatches below class level are for comparisons between sequences of representative ray-finned fishes (class Actinopterygii); for between-class comparisons representative Vertebrata were considered ($n = 155$ sequences; see the electronic supplementary material for details).

samples being compared and sequencing is carried out at sufficient depth. Furthermore, this could be considered a small concession given that COI allows access to a large number of barcode sequences linked to taxonomically verified specimens. However, we feel that even the best COI metabarcoding studies highlight this marker's limitations and indicate that alternatives should be seriously considered. For example, Yu *et al.*'s [2] work on bulk sequencing of COI from arthropod samples for biodiversity analysis documented dropout rates of between 24% (more than 2 read threshold) and 36% (more than 5 read threshold) compared with known inputs even when using fully degenerate primers. While the resultant data produce estimates of α- and β-diversity useful for conservation-relevant decisions [9], acceptance of this level of bias will surely limit future applications. Variation in occurrence of taxa prone to dropout between groups of samples can potentially skew the relative importance of all taxa, making it difficult to assess biologically relevant differences between groups.

When preliminary methodological evaluations are not comprehensive and limitations of the dataset are not taken into account, data interpretation is fraught with difficulties. In a recent study evaluating insect metabarcoding markers [8], a set of widely used 'generic arthropod' COI metabarcoding primers only managed to recover between 43 and 64% of species in a known mixture of arthropod DNA. Retrospective evaluation of ecological studies reliant on data produced from these primers is difficult; however, in some cases primer preferences rather than biology may be driving conclusions.

Increasing sequencing depth to allow detection of poorly amplified markers is unlikely to be a robust solution, because there will be a concomitant increase in the number of sequences originating from minor contamination and chimeric molecules [7,10]. Methods used to filter out these low-level background errors and identify legitimate rare sequences are imperfect. Furthermore, incorporation of low-level errors into metabarcoding datasets can have a disproportionate influence because summaries are typically incidence-based (i.e. presence/absence) and do not include information on sequence abundance.

Despite the large COI reference database being a strong selling point for this marker, many COI metabarcoding studies link recovered sequences to operational taxonomic units (OTU) rather than providing high-resolution taxonomic information [9]. This partly reflects adoption of bioinformatic approaches from microbial ecologists, but it also reflects the lack of coverage within the global COI database. The large collection of COI reference sequences may help improve broad taxonomic assignments (i.e. to family or genus), but in many studies locally developed databases will be required if the intention is to move away from OTU indicators and get back to biology [11]. This opens the possibility of sequencing non-standard barcode markers better suited to metabarcoding when deemed appropriate. Flexibility in which marker is used for metabarcoding is a necessity for some animal groups, such as nematodes, where it is recognized that COI is unsuitable due to sequence diversity. There are also similar issues for 'official' plant barcodes, resulting in many plant metabarcoding studies choosing 'unofficial' markers.

4. What is the way forward?

The accuracy of metabarcoding is highly dependent on marker choice, but there is unfortunately no perfect metabarcoding marker. Instead, the best marker choice is going to be study-specific. For designing highly conserved primers, the mosaic pattern of variation seen in ribosomal RNA (rRNA) genes is often very useful (figure 1). These genes have already been adopted by many in the animal metabarcoding community and are standard markers for fungal and bacterial/archaeal identification. For animals, nuclear rRNA genes

provide very wide taxonomic coverage but lower taxonomic resolution, whereas mitochondrial rRNA genes provide taxonomic resolution similar to COI but typically allow the design of more conserved primers (figure 1). Perceived difficulties in assigning rRNA gene sequences to taxa caused by the inability to accurately align sequences can largely be overcome using alignment-free methods [12]. However, length variation in rRNA coding regions can potentially cause taxon-specific differences in sequence recovery. It is also true that easier alignment of protein genes allows for correction of some sequencing errors [2]. The important point is that a range of potential primers, and the taxonomic resolution of the resulting amplicons, should be carefully considered in any metabarcoding application. The primers can be easily evaluated *in silico* by using available programs (e.g. ecoPCR [3]); empirical testing provides further assurance that primers are suitable for a particular application [2,5,8].

We envisage that metabarcoding will eventually routinely sequence several barcode markers from each sample [10,13]. Markers aimed at different taxonomic levels can overcome the trade-off between taxonomic breadth and resolution. Markers providing comparable taxonomic information can act as internal controls; these would be especially useful for validation in cases where primer–template mismatches are a potential problem. Metabarcoding approaches relying on bulk sequencing of enriched mtDNA without amplification have been illustrated in a proof of concept study [14]. This work may well point to a future where PCR primers are less relevant; however, methods outlined so far require intact mtDNA molecules and would not be applicable when DNA is highly fragmented. Alternative marker-enrichment techniques that work with a range of templates, such as probe capture-based approaches, might be better suited to non-COI markers that contain conserved target regions.

We acknowledge that there are situations where COI could currently be the preferred option as a metabarcoding marker (e.g. when taxonomic scope is limited and species-level identification critical, or when the existing reference database is essential). Indeed, if future techniques allow less-biased recovery of COI from DNA mixtures, COI would be well suited to metabarcoding. Even if alternative markers are adopted, the DNA barcoding infrastructure developed by CBOL will be vital for this field. Taxonomically verified voucher specimens, and associated DNA extracts, are an invaluable resource that could facilitate high-throughput characterization of additional markers [15]. The CBOL database with reference sequences linked to voucher specimens (including 'unofficial' barcode sequences), and efforts to link CBOL's taxonomic metadata to publicly accessible sequences in GenBank, are equally beneficial. We are excited by the prospect of metabarcoding providing a faster and less expensive method to measure animal biodiversity, but marker selection needs more scrutiny and available marker choices need to be broadened for improved reliability.

Data accessibility. The DNA sequences extracted from GenBank and used for construction of figures 1 and 2 are deposited as electronic supplementary data.

Acknowledgements. We thank our colleagues for discussions on this topic. We also thank the three reviewers for providing critical comments that helped improve the manuscript.

Funding statement. B.D. and S.J. received operating grants from the Australian Antarctic Science Program (AAS Projects 4014 and 4313).

References

1. Taberlet P, Coissac E, Hajibabaei M, Rieseberg LH. 2012 Environmental DNA. *Mol. Ecol.* **21**, 1789–1793. (doi:10.1111/j.1365-294X.2012.05542.x)
2. Yu DW, Ji Y, Emerson BC, Wang X, Ye C, Yang C, Ding Z. 2012 Biodiversity soup: metabarcoding of arthropods for rapid biodiversity assessment and biomonitoring. *Methods Ecol. Evol.* **3**, 613–623. (doi:10.1111/j.2041-210X.2012.00198.x)
3. Ficetola GF, Coissac E, Zundel S, Riaz T, Shehzad W, Bessiere J, Taberlet P, Pompanon F. 2010 An *in silico* approach for the evaluation of DNA barcodes. *BMC Genomics* **11**, e434. (doi:10.1186/1471-2164-11-434)
4. Geller J, Meyer C, Parker M, Hawk H. 2013 Redesign of PCR primers for mitochondrial cytochrome *c* oxidase subunit I for marine invertebrates and application in all-taxa biotic surveys. *Mol. Ecol. Resour.* **13**, 851–861. (doi:10.1111/1755-0998.12138)
5. Klindworth A, Pruesse E, Schweer T, Peplies J, Quast C, Horn M, Glockner FO. 2013 Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies. *Nucleic Acids Res.* **41**, e1. (doi:10.1093/nar/gks808)
6. 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. *Appl. Environ. Microbiol.* **74**, 1660–1663. (doi:10.1128/aem.02403-07)
7. Schloss PD, Gevers D, Westcott SL. 2011 Reducing the effects of PCR amplification and sequencing artifacts on 16S rRNA-based studies. *PLoS ONE* **6**, e27310. (doi:10.1371/journal.pone.0027310)
8. Clarke LJ, Soubrier J, Weyrich LS, Cooper A. In press. Environmental metabarcodes for insects: *in silico* PCR reveals potential for taxonomic bias. *Mol. Ecol. Resour*. (doi:10.1111/1755-0998.12265)
9. Ji Y *et al*. 2013 Reliable, verifiable and efficient monitoring of biodiversity via metabarcoding. *Ecol. Lett.* **16**, 1245–1257. (doi:10.1111/ele.12162)
10. De Barba M, Miquel C, Boyer F, Mercier C, Rioux D, Coissac E, Taberlet P. 2014 DNA metabarcoding multiplexing and validation of data accuracy for diet assessment: application to omnivorous diet. *Mol. Ecol. Resour.* **14**, 306–323. (doi:10.1111/1755-0998.12188)
11. Leray M, Yang JY, Meyer CP, Mills SC, Agudelo N, Ranwez V, Boehm JT, Machida RJ. 2013 A new versatile primer set targeting a short fragment of the mitochondrial COI region for metabarcoding metazoan diversity: application for characterizing coral reef fish gut contents. *Front. Zool.* **10**, e34. (doi:10.1186/1742-9994-10-34)
12. Little DP. 2011 DNA barcode sequence identification incorporating taxonomic hierarchy and within taxon variability. *PLoS ONE* **6**, e20552. (doi:10.1371/journal.pone.0020552)
13. Deagle BE, Kirkwood R, Jarman SN. 2009 Analysis of Australian fur seal diet by pyrosequencing prey DNA in faeces. *Mol. Ecol.* **18**, 2022–2038. (doi:10.1111/j.1365-294X.2009.04158.x)
14. Zhou X *et al*. 2013 Ultra-deep sequencing enables high-fidelity recovery of biodiversity for bulk arthropod samples without PCR amplification. *GigaScience* **2**, 4. (doi:10.1186/2047-217X-2-4)
15. Shokralla S, Gibson JF, Nikbakht H, Janzen DH, Hallwachs W, Hajibabaei M. 2014 Next-generation DNA barcoding: using next-generation sequencing to enhance and accelerate DNA barcode capture from single specimens. *Mol. Ecol. Resour.* **14**, 892–901. (doi:10.1111/1755-0998.12236)