NOTE / NOTE

Benchmarking DNA barcodes: an assessment using available primate sequences

Mehrdad Hajibabaei, Gregory A.C. Singer, and Donal A. Hickey

Abstract: DNA barcoding has been recently promoted as a method for both assigning specimens to known species and for discovering new and cryptic species. Here we test both the potential and the limitations of DNA barcodes by analysing a group of well-studied organisms—the primates. Our results show that DNA barcodes provide enough information to efficiently identify and delineate primate species, but that they cannot reliably uncover many of the deeper phylogenetic relationships. Our conclusion is that these short DNA sequences do not contain enough information to build reliable molecular phylogenies or define new species, but that they can provide efficient sequence tags for assigning unknown specimens to known species. As such, DNA barcoding provides enormous potential for use in global biodiversity studies.

Key words: DNA barcoding, species identification, primate, biodiversity.

Résumé : L'emploi de codes-barre a récemment été proposé pour assigner des spécimens à des espèces connues et pour découvrir des espèces nouvelles et cryptiques. Dans ce travail, les auteurs testent le potentiel et les limitations des codes-barre d'ADN en analysant un group d'organismes bien étudiés : les primates. Les résultats montrent que les codes-barre fournissent suffisamment d'information pour identifier et délimiter efficacement les espèces de primates, mais ils ne permettent pas de révéler plusieurs des relations phylogénétiques plus fines. Les auteurs concluent que ces courtes séquences d'ADN ne sont pas suffisamment informatives pour construire des phylogénies moléculaires fiables ou pour définir de nouvelles espèces. Cependant, ils constituent des « étiquettes » moléculaires efficaces pour assigner un spécimen inconnu à une espèce connue. À ce titre, les codes-barre sont d'une grande utilité potentielle pour l'étude de la biodiversité globale.

Mots clés : code-barre d'ADN, identification des espèces, primate, biodiversité.

[Traduit par la Rédaction]

DNA barcodes have been proposed as a powerful new method for quickly identifying known species, discovering unknown species, and pinpointing cryptic species (Blaxter 2003; Hebert et al. 2003; Marshall 2005). This technique has also been the subject of some criticism (Moritz and Cicero 2004; Ebach and Holdrege 2005; Marshall 2005), mainly that short barcode sequences do not constitute an adequate species description and that they do not contain enough information to infer species relationships. Here we perform a simple test to demonstrate both the potential and the limits of the barcoding approach. Essentially, we agree with the concerns of other authors concerning the use of short barcode sequences, either as a basis for phylogenetic classification or for the definition of cryptic species. Nevertheless, we have shown that DNA barcodes provide a very efficient way of assigning individuals to known species.

We chose a group of well-studied organisms—the primates—for which both the species boundaries and the species relationships are established. We downloaded the available primate barcode sequences from GenBank (see Supplementary Table[2]

Received 6 September 2005. Accepted 12 February 2006. Published on the NRC Research Press Web site at http://genome.nrc.ca on 21 July 2006.

Corresponding Editor: B. Golding.

M. Hajibabaei. Biodiversity Institute of Ontario, Department of Integrative Biology, University of Guelph, Guelph, ON N1G 2W1, Canada.
G.A.C. Singer. Human Cancer Genetics Program, The Ohio State University, Columbus, OH 43210, USA.
D.A. Hickey.[1] Department of Biology, Concordia University, 7141 Sherbrooke Street, Montreal, QC H4B 1R6, Canada.

[1]Corresponding author (e-mail: dhickey@alcor.concordia.ca).

[2]Supplementary data for this article are available on the journal Web site (http://genome.nrc.ca) or may be purchased from the Depository of Unpublished Data, Document Delivery, CISTI, National Research Council Canada, Building M-55, 1200 Montreal Road, Ottawa, ON K1A 0R6, Canada. DUD 5066. For more information on obtaining material refer to http://cisti-icist.nrc-cnrc.gc.ca/irm/unpub_e.shtml.

 doi:10.1139/G06-025

Fig. 1. DNA barcoding of 703 sequences from 28 primate species based on mitochondrial gene cytochrome *c* oxidase subunit I (*COI*) sequences (the full list of GenBank accession numbers is provided in the Supplementary Table[2]). Shown is a neighbor-joining tree made from a 651 bp sequence of the *COI* gene.

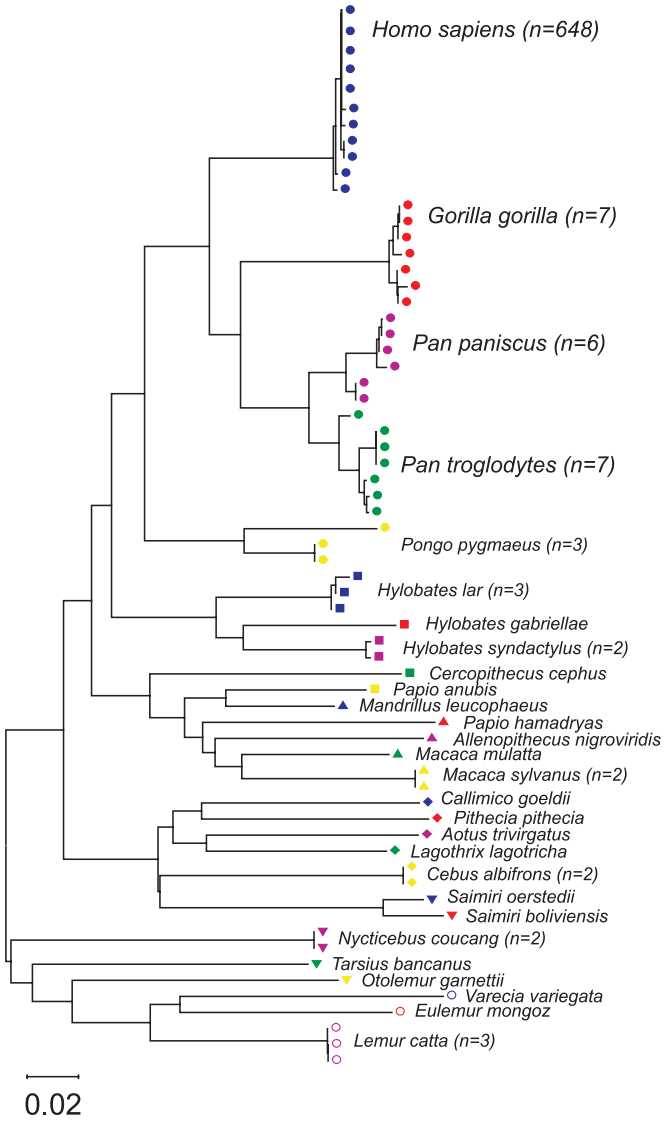
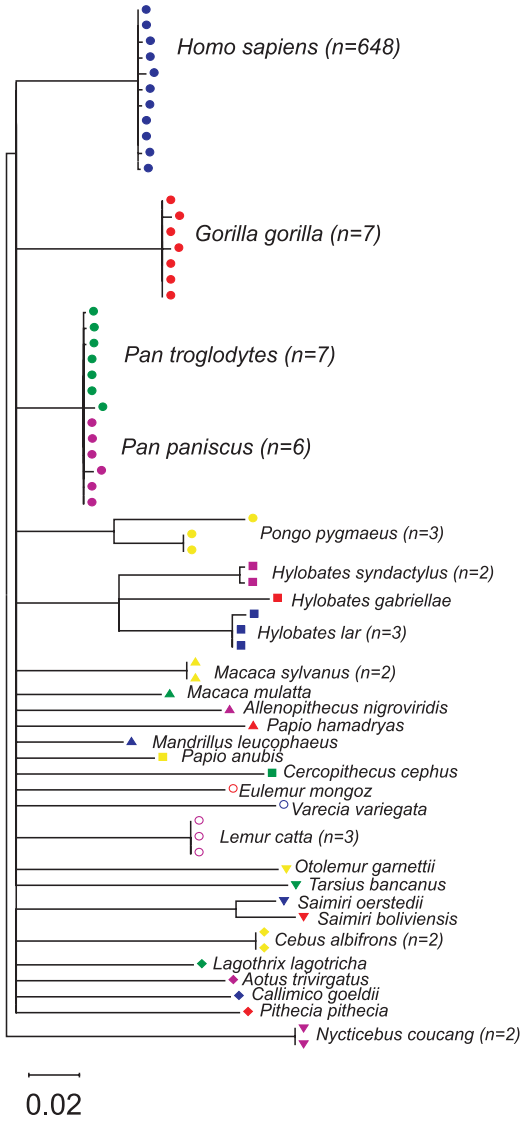


Fig. 2. The tree from Fig. 1, with a bootstrap analysis in which branches with less than 100% support are collapsed.



for list of sequence accession numbers). Our goal was to test the reliability of these sequences for 2 different, but related purposes. First, we wanted to ask if all sequences from a given species would be grouped together. Secondly, we wished to know if the higher level groupings would be statistically well supported and if they would match the accepted phylogeny of primate species.

All sequences were aligned manually, using the reading frame as a guide. Based on these aligned sequences, neighbor-joining trees were constructed according to the methods of Saitou and Nei (1987) using the Mega3 software package (Kumar et al. 2004). We chose the Kimura model of base substitution (Kimura 1980) and performed 1000 replicates for bootstrapping analysis (Felsenstein 1985). The results are shown in Fig. 1.

The first goal of the present study, as stated above, was to see if all of the sequences from the same species clustered together in the neighbor-joining tree. Figure 1 illustrates that this is indeed the case; barcode sequences from a given species always group together in the tree. In contrast to this, many of the higher order branches are poorly supported and do not match the accepted phylogenetic relationships within the primates. This is reflected both in the low bootstrap values for many of the internal nodes and the inability to resolve short branches correctly. For example, the placement of humans as an outgroup to chimps and gorillas is incorrect. In summary, several of the groupings above the species level are incorrect and many of them are not statistically well supported.

To filter out the poorly supported parts of the tree, we collapsed all branches that did not have 100% bootstrap support (Fig. 2). This second tree retained the species identification at the tips of the tree (with the exception of lumping the 2 closely related species of *Pan*), while removing many of the poorly supported internal nodes. To put it simply, DNA barcodes may delineate individual species with a high degree of

confidence but they do not reliably uncover many of the deeper phylogenetic relationships.

The one exception to unique species identification at the very high (100%) bootstrap value was the failure to resolve the barcodes for the 2 species of chimpanzee, *Pan paniscus* and *Pan troglodytes* (see Fig. 2). When the cut-off is lowered to a 95% bootstrap value, the 2 species of chimp are resolved (Fig. 3). It is interesting to note that the human–chimp–gorilla trichotomy is still not resolved in this figure. An alternative to relaxing the bootstrap cut-off value is to retain the 100% value and increasing the length of the sequence used for the barcode. Therefore, we extended the barcode sequence to 1500 bp and, as expected, the use of longer sequences resulted in the resolution of the 2 species of chimp, even at the higher bootstrap cut-off value (Fig. 4). This result highlights the fact that there is no single, "correct" length for a DNA barcode. Although longer sequences give greater resolution, unexpectedly short sequences provide excellent resolution at the species level. Thus it is far more efficient to use very short sequences for the initial screening of large numbers of samples. Difficult cases, owing to recent divergences and (or) low rates of molecular evolution, can be resolved by extending the barcode sequence. Of course, more challenging problems such as molecular phylogenetics or DNA-based species definitions would require far more extensive sequences than those that are needed for species barcoding.

This analysis is not meant to be exhaustive, but to highlight 2 important features of DNA barcoding. First, despite the short sequences used, this method provides an efficient means of generating "molecular labels" for species. Second, and precisely because of the short sequence length, DNA barcoding cannot reliably quantify the relationships between species and therefore should not be used to build deep molecular phylogenies. Consequently, we agree with those who claim that DNA barcoding is not a threat to either taxonomy or biosystematics (Schindel and Miller 2005).

It should be noted that although we have shown that DNA barcoding can accurately assign samples to well-defined primate species, this does not mean that it will be equally applicable to other groups. But it is does mean that it can function very well in a case where there are no underlying taxonomic difficulties. In fact, it is interesting to note that the 2 species of chimpanzee that were difficult to resolve by traditional taxonomic methods also provide a challenge for the barcoding approach. It remains to be seen how DNA barcoding will work for other species groups, especially in groups that lack outbreeding or that are characterized by significant amounts of geographic differentiation and (or) hybridization. It is worth noting that the rates of divergence for COI sequences in primates (an average of 0.30% for conspecific comparisons and 5.88% for congeneric comparisons) are comparable with what is found among the barcode sequences of fish (0.39% and 9.93%; Ward et al. 2005) and birds (0.43% and 7.93%; Hebert et al. 2004).

In summary, we have demonstrated that short barcoding sequences contain sufficient information for the reliable delineation of species, but not enough for the assembly of a complete phylogenetic tree. Thus, barcoding provides a checklist for the "leaves" of the phylogenetic tree rather than a map of the branching pattern (Crandall and Buhay 2004);

Fig. 3. The hominid portion of tree in Fig. 1 in which nodes with a bootstrap value less than 95% are collapsed.

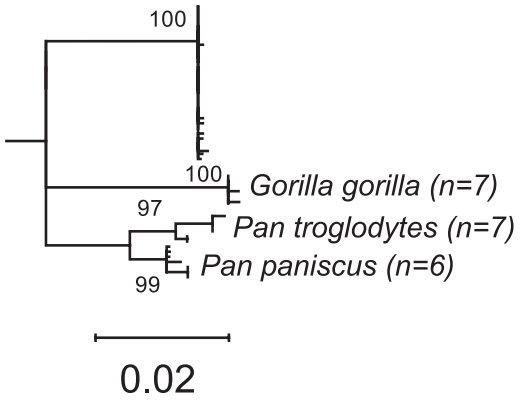
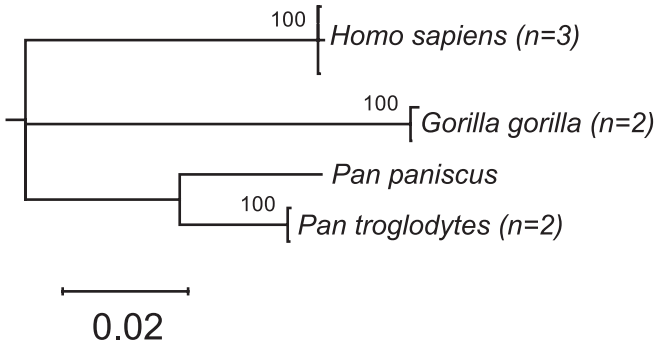


Fig. 4. A DNA barcoding tree for the hominid species with a 1500 bp (full length) *COI* gene. Note that, owing to space constraints, only a representative subset of the human sequences are shown. The complete Figure can be accessed as Supplementary Fig. 1.[2] However, these data points do not show any significant variation from the presented set.



as such, DNA barcoding has enormous potential for use in global biodiversity studies. It is a tool that enables us to perform high-throughput analyses of species abundances on a global scale and to track changes in those abundances through time. Barcoding is ideal for the field biologist who wants to assess species-level biodiversity in a large geographic area in a short amount of time, since the process is quick and can be automated. The pre-sorted samples could then be passed on to an expert taxonomist for a more detailed analysis, if required.

References

Blaxter, M. 2003. Molecular systematics: counting angels with DNA. Nature (London), **421**: 122–124.

Crandall, K.A., and Buhay, J.E. 2004. Evolution. Genomic databases and the tree of life. Science (Washington, D.C.), **306**: 1144–1145.

Ebach, M.C., and Holdrege, C. 2005. DNA barcoding is no substitute for taxonomy. Nature (London), **434**: 697.

Felsenstein, J. 1985. Confidence limits on phylogenies: an approach using the bootstrap. Evolution, **39**: 783–791.

Hebert, P.D.N., Cywinska, A., Ball, S.L., and deWaard, J.R. 2003. Biological identifications through DNA barcodes. Proc. Biol. Sci. **270**: 313–321.

Hebert, P.D.N., Stoeckle, M.Y., Zemlak, T.S., and Francis, C.M. 2004. Identification of birds through DNA barcodes. PLoS Biol. **2**: E312.

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

Kumar, S., Tamura, K., and Nei, M. 2004. MEGA3: Integrated software for Molecular Evolutionary Genetics Analysis and sequence alignment. Brief Bioinform. **5**: 150–163.

Marshall, E. 2005. Taxonomy. Will DNA bar codes breathe life into classification? Science (Washington, D.C.), **307**: 1037.

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

Saitou, N., and Nei, M. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. **4**: 406–425.

Schindel, D.E., and Miller, S.E. 2005. DNA barcoding a useful tool for taxonomists. Nature (London), **435**: 17.

Ward, R.D., Zemlak, T.S., Innes, B.H., Last, P.R., and Hebert, P.D.N. 2005. DNA barcoding Australia's fish species. Philos Trans R. Soc. Lond. B Biol. Sci. **360**: 1847–1857.