Non-COI Barcode Regions — Guidelines for CBOL Approval

The Consortium for the Barcode of Life (CBOL) has so far accepted the 648 base-pair 'Folmer region' of COI (mitochondrial encoded *cytochrome oxidase* 1) as the default DNA barcode region for vertebrates and insects and promotes its use in as many other clades as possible. This widespread adoption of COI as barcode region offers significant benefits to researchers and barcode users. The International Nucleotide Sequence Database Collaboration (INSDC, consisting of GenBank, the European Molecular Biology Laboratory and the DNA Data Bank of Japan) has adopted the data standards proposed by CBOL for BARCODE data records, and has empowered CBOL to decide which gene regions can be given BARCODE status.

In addition to promoting standardization of barcode regions, CBOL also seeks extending the application of DNA barcoding across all eukaryotic life. CBOL recognizes that:

- COI does not vary in some taxonomic groups, or is prone to exceptional molecular evolutionary processes in ways that prevent it from being an effective barcode region;
- COI may not be able to resolve species-level differences in all subgroups of a taxonomic group, and additional sequence data may be needed from a second or even third region in such cases; and
- Researchers may have already gathered significant volumes of data using a different gene region in a particular taxonomic group which would, if properly vouchered, provide good potential as a DNA barcode region.

For these reasons, CBOL has developed guidelines for the adoption of gene regions other than COI as the barcode region for a particular clade. These guidelines specify the documentation that must be submitted to CBOL as an application for BARCODE status in INSDC. Proposals will be reviewed by CBOL's Scientific Advisory Board, which will provide its advice to CBOL's Executive Committee, which will reach a final decision.

In order to support adopting non-COI regions for DNA barcoding the following questions need to be addressed:

a) Has COI been proven ineffective as a barcode region for the clade under consideration?
b) Have alternative candidate regions been tested, screened, and compared in a systematic manner, i.e., can they distinguish sisters in (as many as possible) species pairs within the clade?
c) Does the proposed gene region work effectively as a barcode across the clade under consideration, and is it easily PCR amplified even from degraded tissues?
d) Are a few 'universal' primers likely to be effective, or will extensive primer development be needed?
e) There may be clade-specific biological reasons for adopting an alternative method of assigning barcode regions. What is the rationale and performance of the proposed method?

In order to standardize the assessment of non-COI regions as much as possible we ask proposers to use the following guidelines when preparing their documentation.

1. **Rejection of COI.** Before CBOL will consider a non-COI region, applicants must document the ineffectiveness of COI as a barcode region in the taxonomic group of interest. Proposers must provide evidence that addresses the following:

   a) ***PCR problems***. In cases where COI is rejected as a barcode region owing to the inability to extract or amplify COI, the proposer must document his/her efforts to:

      i) test different extraction methods and amplification protocols;
      ii) develop new primers; and
      iii) consult with other barcode researchers.

   b) ***Pattern of intra and inter specific variation***. Using data from INSDC, or other public databases, proposers should collect sequences of the COI from several groups of sibling species across the clade of interest. Sequence divergence *d* should be estimated with a model of evolution appropriate at that level of variation (e.g. K2P). Intra- and interspecific variation should be compared and plotted as in Fig. 1, and an associated COI NJ tree should be provided as well in order to demonstrate the failure of specific clustering. Finally, if applicable, a description of possible variation due to the presence of indels should be included.

   c) ***Resolving power***. Using data from INSDC or other public databases, proposers should document the incapacity of COI to discriminate between as many sibling species pair as possible, with each species represented by multiple individuals from different geographical areas of its range. Individuals should be sampled following the criteria already established by CBOL's Data Base Working Group (clear sample locations, taxonomic identifications, availability of vouchers).

2. **Selection of non-COI barcode region.** Proposers must document the process used to identify the proposed region as the optimal barcode region for the taxonomic group under consideration. CBOL will not accept a proposal based only on the volume of sequence data that has been collected for a gene region in the past. What other candidate regions have been tested? How were they screened and compared? In any case divergence within and among species should be shown using K2P distances to test their effectiveness as barcodes.

3. **Performance of the Non-COI Barcode Region and /or alternative selection method.** Proposers must present the following evidence that the proposed region works effectively as a barcode across the taxonomic group under consideration (see Table 1):

   a) ***Pattern of intra- and interspecific variation***. Using data from INSDC, or other public databases, proposers should collect sequences of the new marker from several groups of sibling species across the clade of interest. Sequence divergence *d* should be estimated with a model of evolution appropriate at that level of variation (e.g., K2P). Intra- and interspecific variation of the proposed non-COI region should be compared and plotted as in Fig. 1. Intraspecific divergence should be calculated as the arithmetic mean of pairwise divergence within a

species, using that model. Interspecific divergence between sister-species pairs should be calculated as their inter-centroid distances. Ideally the presence of a 'barcode gap' (i.e., interspecific divergences being greater than intraspecific divergence) would emerge from this, however this is not an absolute requirement for approval. In case of absence of a barcode gap, proposers should provide a Neighbor-Joining tree of the sequence distances as well in order to demonstrate specific clustering. In any case, a description of possible variation due to the presence of indels should be included.

b) ***Resolving power***. Using data from INSDC or other public databases, proposers should document the capacity of the new marker and the chosen method to discriminate between as many sibling species pairs as possible, with each species represented by multiple individuals from different geographical areas of the species range. Individuals should be sampled following the criteria already established by CBOL's Data Base Working Group (clear sample locations, taxonomic identifications, availability of vouchers).

c) ***Universality***. The effectiveness of the best primer pair for the proposed region should be documented, along with the return on investment in developing extra customized primers for generating barcodes in the targeted clade. This information should demonstrate if it will be possible and worthwhile developing truly universal primers for the clade (see also Table 1).

Implementation. The current protocols will be adopted for a period of 6 months during which CBOL is open to suggestions for their improvement from the community. If, for instance, alternative methods of selection the barcode region are proposed, the rationale for doing so should be documented. For example, different clades could have different reproductive, life history, and/or molecular evolutionary features such that some approaches are more effective than others. Proposers should document their analysis as described in point 3, above.

CBOL will normally expect proposers to have presented evidence of the effectiveness of their proposed non-COI barcode region(s) in a peer-reviewed publication prior to submission of a proposal. Prior peer review and publication will support the proposal's claims and will inform the community of the proposed barcode region(s).

After receiving a proposal for a non-COI barcode region and any supporting reprints or preprints, CBOL will evaluate the proposal in consultation with the scientific community. If a proposal is approved by CBOL's Executive Committee, INSDC will be informed immediately of the decision to give BARCODE status to the proposed region(s).

Recommended for adoption by CBOL Scientific Advisory Board, March 2007

Approved by CBOL Executive Committee, 14 May 2007

Fig 1. *DNA barcode gap.* Sequence divergence for different gene regions across three species. Intraspecific divergence is calculated as the arithmetic mean of pairwise divergence within a species, using an appropriate model for sequence divergence. Interspecific divergence between species pairs is calculated as their inter-centroid distances.

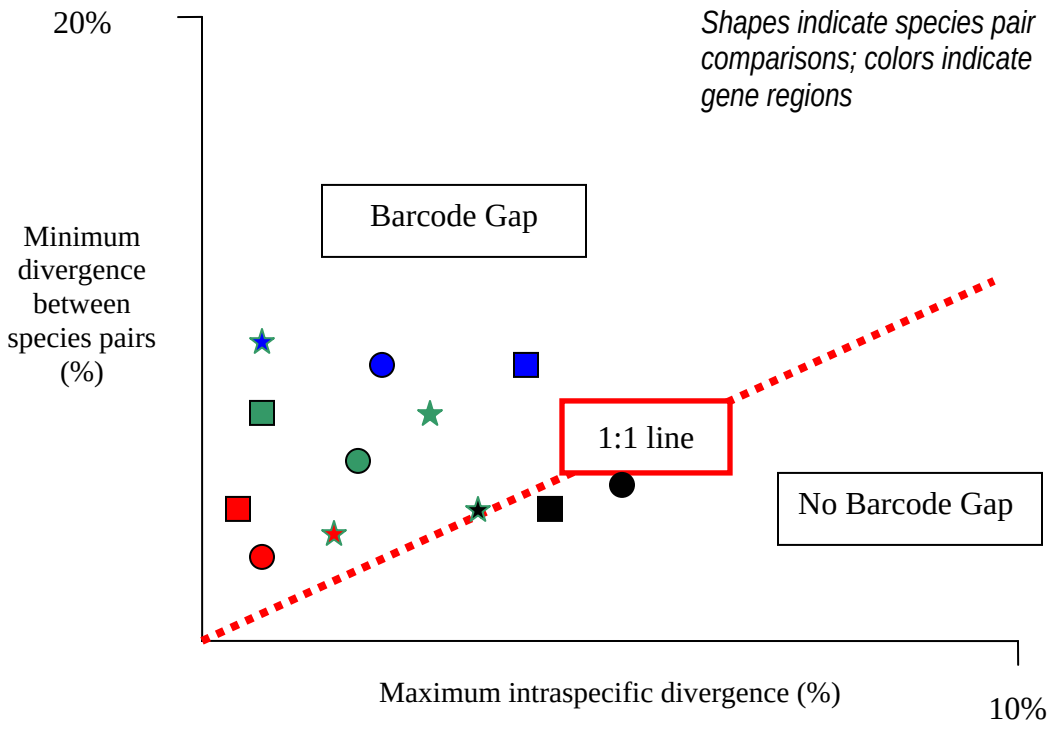

Table 1. Summary presented separately for each major subclade as well as for the entire clade. If multiple barcode regions are being proposed, then a separate column must be included and filled in for each proposed gene region A, B, C and D (four columns are provided below only as an example).

	Region A	Region B	Region C	Region D
Number of species pairs tested[1]				
% of species successfully amplified (please note in parentheses the number of primer pairs used)	% success (no. primer pairs)	% success (no. primer pairs)	% success (no. primer pairs)	% success (no. primer pairs)
% of species successfully identified using single best region[2]				
% of species successfully identified using all proposed regions together				
Range of intraspecific variation[3]	Min-Max	Min-Max	Min-Max	Min-Max
Median % of intraspecific variation[3]				
Range of divergence between sister species pairs[4]	Min-Max	Min-Max	Min-Max	Min-Max
Median % of minimum divergences between sister species pairs[4]				

[1]Please indicate the number of specimens tested per species

[2]Results of cluster analysis must be provided

[3]Based on average appropriate DNA sequence distance across multiple specimens representing the biogeographical range of the species

[4]Proposers must specify the method used to calculate divergence between sister species pairs (centroids, minimum distance, etc.)