MOLECULAR ECOLOGY

Molecular Ecology (2016) **25**, 1908–1910

NEWS AND VIEWS

PERSPECTIVE

Coalescing molecular evolution and DNA barcoding

LUCIE ZINGER* and HERVÉ PHILIPPE†
*CNRS, ENFA, UMR 5174 EDB, Université Toulouse 3 Paul Sabatier, F-31062 Toulouse, France; †Centre de Théorisation et de Modélisation de la Biodiversité, UMR CNRS 5321, Station d'Ecologie Théorique et Expérimentale, Moulis 09200, France

The DNA barcoding concept (Woese *et al.* 1990; Hebert *et al.* 2003) has considerably boosted taxonomy research by facilitating the identification of specimens and discovery of new species. Used alone or in combination with DNA metabarcoding on environmental samples (Taberlet *et al.* 2012), the approach is becoming a standard for basic and applied research in ecology, evolution and conservation across taxa, communities and ecosystems (Scheffers *et al.* 2012; Kress *et al.* 2015). However, DNA barcoding suffers from several shortcomings that still remain overlooked, especially when it comes to species delineation (Collins & Cruickshank 2012). In this issue of *Molecular Ecology*, Barley & Thomson (2016) demonstrate that the choice of models of sequence evolution has substantial impacts on inferred genetic distances, with a propensity of the widely used Kimura 2-parameter model to lead to underestimated species richness. While DNA barcoding has been and will continue to be a powerful tool for specimen identification and preliminary taxonomic sorting, this work calls for a systematic assessment of substitution models fit on barcoding data used for species delineation and reopens the debate on the limitation of this approach.

Keywords: clustering methods, genetic distances, posterior predictive simulations, species delimitation

Received 16 March 2016; revision received 31 March 2016; accepted 31 March 2016

The establishment of gold-standard genomic regions (hereafter DNA barcodes) such as the ribosomal rRNA gene regions for microorganisms, the mitochondrial cytochrome c oxidase (CO1) for animals or the chloroplastic large subunit of ribulose-1,5-biphosphate carboxylase for plants (*rbcL*, reviewed in Taberlet *et al.* 2012) has revolutionized taxonomy research by rendering possible an approximate classification of most kinds of organisms in a standard and time-effective way. The approach has a resolution that was *hitherto* unreachable: while morphospecies are discriminated on the basis of tens of characters that may be sometimes subjective or invisible in cryptic species, a barcode contains hundreds of unambiguous characters. Thanks to next-generation sequencing, it is now possible to produce barcodes from thousands of specimens or environmental samples (e.g. soil or water, Taberlet *et al.* 2012). Hence, DNA (meta)barcoding allows to rapidly characterize the species composition and richness of communities at unprecedented spatial/temporal scales and taxonomic breadth. This opens promising perspectives in biodiversity monitoring and management, for example for identifying areas of high conservation value or threatened by anthropogenic activities. More fundamentally, it unequivocally improves our understanding of the processes that maintain biodiversity, which is crucial if we are to better estimate extinctions rates in the current era of biodiversity loss (Scheffers *et al.* 2012; Kress *et al.* 2015).

Species can be defined as lineages that evolve separately from each other (De Queiroz 2007). The DNA barcoding concept relies on the idea that they can be differentiated because barcode interspecific genetic distances exceed intraspecific distances (Hebert *et al.* 2004; Puillandre *et al.* 2012). When used to affiliate specimens or sequences to barcoded species, the approach is usually appropriate and straightforward. The specimen is assigned to a given species if its sequence identity (defined without accurate models of sequence evolution, Collins & Cruickshank 2012) to a reference barcode is above a given threshold (usually 99% of sequence identity, Ratnasingham & Hebert 2007). Identification success heavily relies on the comprehensiveness of existing barcode databases (e.g. http://www.barcodinglife.org) and their accuracy, but risk of misidentification is relatively limited for many plants and animals. However, reference databases, albeit rapidly growing, remain far from complete in most eukaryotic groups with high taxonomic impediments (e.g. microfauna or microorganisms). A common approach to circumvent this problem is to delineate species from barcoding data in an unsupervised way, that is based on genetic variation. But this approach has been repeatedly questioned (Collins & Cruickshank 2012; Dupuis *et al.* 2012). First, because DNA barcoding studies often consider only one locus, which precludes taking into account the full evolutionary history of species (in particular introgression and hybridization). Second because partitioning the genetic variation between species or specimens constitutes a mathematical challenge on its own. This has led to the development of numerous partition methods to form operational taxonomic units (OTUs), which are usually considered as putative species. These methods are based either on pairwise genetic distances (e.g. ABGD, BIN, UPARSE), or on time-calibrated ultrametric or

Correspondence: Hervé Philippe, Fax: +33 561 960 851;
E-mail: herve.philippe@ecoex-moulis.cnrs.fr

nonultrametric phylogenies (e.g. GMYC, PTP), each with their own *pros* and *cons* (see Coissac *et al.* 2012 for a review).

Hence, genetic distances are central to the barcoding concept, and should be generated with an appropriate model of sequence evolution. This issue has been the focus of active research in phylogenetics/phylogenomics (Philippe *et al.* 2011a), but has been comparatively seldom considered by the barcoding community (Collins & Cruickshank 2012). In this issue of *Molecular Ecology*, Barley & Thomson (2016) bridge this gap by assessing substitution model adequacy in describing genetic variation and estimating species richness from barcoding data. The work is mainly motivated by the wide use of the simplistic Kimura 2-parameter model (K2P, Kimura 1980) for generating distance matrices, although its underlying assumptions are most likely violated by barcoding data (Galtier *et al.* 2009). For instance, the K2P model assumes that the equilibrium frequencies of the 4 nucleotides are equal, whereas mitochondrial genomes are generally AT-rich. It also assumes that the same substitution process applies to the first, second or third codon position.

To evaluate substitution model adequacy, Barley & Thomson (2016) used posterior predictive simulations (Fig. 1, see Barley & Thomson 2016 for more details). The analysis of posterior predictive distributions is widely used in Bayesian statistics to assess model plausibility, but remains underused in phylogenetics or in ecology (but see e.g. Brown 2014, Kery & Royle 2015). Briefly, it is similar to the parametric bootstrapping of maximum-likelihood statistics, except that it integrates parameter estimation uncertainties. Its principle consists in (i) simulating data according to the model and data under study and (ii) comparing real and simulated data or any type of inference based on them (Fig. 1). The more similar the real and simulated data are, the better the model is. The key to this approach is to invent clever statistics that will allow answering questions of interest.

Using posterior prediction and data-based test statistics, Barley & Thomson (2016) demonstrate that the K2P model fits poorly to DNA barcoding data sets especially when data set complexity is high (e.g. diversified groups such as arthropods), in agreement with previous observations based on model selection procedures (Collins & Cruickshank 2012). But a poorly fitting model can still perform well to infer some parameters. This is the case for phylogenetic trees under some circumstances (Yang 1997), for instance because of a more reduced variance. Accordingly, Barley & Thomson (2016) used the number of OTUs as the inference-based test statistic, which were inferred with two popular clustering methods and different genetic divergence thresholds. They convincingly demonstrated that the K2P model underestimates species richness (Fig. 1).

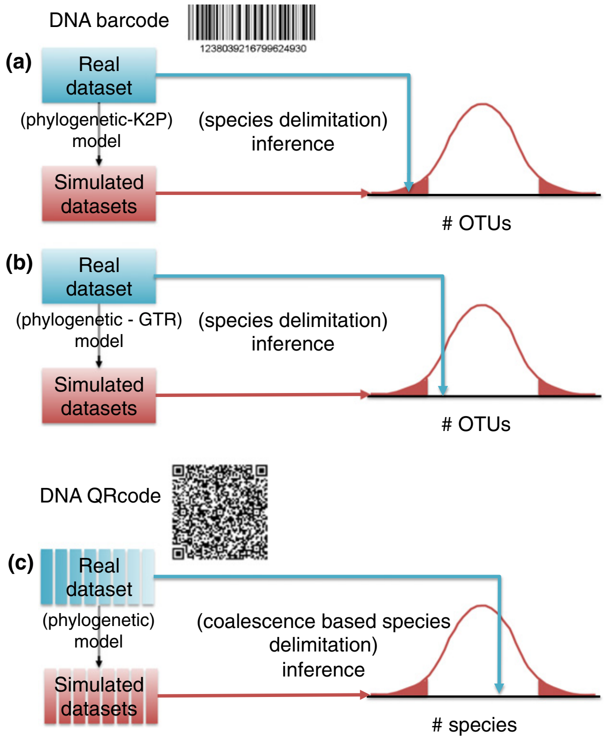

Fig. 1 Model adequacy assessment using post-predictive simulations. The approach consists in inferring the distribution of parameters (posterior) of a given model, simulating new data sets (posterior predictive) using parameters values selected from the posterior distribution and comparing the observed data to the posterior predictive data, either directly using data-based test statistics (not shown here) or inference-based test statistics (here the number of OTUs). (a) The use of the K2P model tends to underestimate OTU richness while more complex models (b) perform better. Using a multilocus approach will undoubtedly refine estimates of species richness (c).

The study of Barley & Thomson (2016) is timely in an era where DNA (meta)barcoding is becoming the tool of choice for sensing biodiversity. It emphasizes the importance of choosing more complex models (e.g. GTR + Γ) that better fit the data if we are to obtain reliable estimates of species richness. This has major implications because high-priority conservation areas and inferences of species extinction rates are based on species richness. The increasing size of barcoding data sets offers the opportunity to use more complex (and hopefully more realistic) models already developed in phylogenetics. In particular, in barcoding data, a nucleotide difference can reflect new neutral mutations or mutations that have been fixed (i.e. substitutions). New models should hence separate mutation rates and fixation probabilities (see De Maio *et al.* 2013). However, these models are computationally demanding, which is inadequate with the increasing size of barcoding data sets. Still, model fit and model parameter values inferred from a subset of randomly selected sequences will likely be accurate. The computational burden of estimating evolutionary distances will not increase substantially when parameter values are fixed. Model selection should hence be applied to species delineation based on barcoding data.

Using the best substitution model does not alleviate the limits of single locus approaches for delineating species.

Following the lineage species concept (De Queiroz 2007), this requires having access to the full evolutionary history of species, which results from multiple mechanisms (e.g. disrupted gene flow, local adaptation, random drift, introgression). This cannot be captured with single standard barcodes (Dupuis *et al.* 2012). New directions such as genome skimming (Coissac *et al.* 2016) will undoubtedly refine our ability to delineate species by giving access to more informative sites in at least 2 or 3 loci (organelle genomes and ribosomal genes for eukaryotes, as well as single copy genes in certain cases). With the decreasing cost of high-throughput sequencing, we foresee an extension of the DNA barcode concept (1Dimension) to the DNA QRcode (2Dimensions, Fig. 1) based on full (meta)genomes. This will constitute a formidable step forward to identify species in all their historical complexities and allow delineating evolutionary significant units, that is not only species but also populations, both being relevant for conservation purposes. It would require using coalescence- rather than phylogeny-based models, and the posterior predictive approach used by Barley & Thomson (2016) will be key to validate these new approaches. However, these advances will also come along with new needs and challenges in bioinformatics and computational biology to adequately handle the newly generated trillions of nucleotides. The environmental footprint of these fascinating scientific developments will be enormous, and it is unclear whether the perspective of improvements in biodiversity conservation will compensate the certain destruction that they will bring along (see Philippe 2011b for a more complete discussion).

References

Barley A, Thomson R (2016) Assessing the performance of DNA barcoding using posterior predictive simulations. *Molecular Ecology*, **25**, 1930–1943.

Brown JM (2014) Detection of implausible phylogenetic inferences using posterior predictive assessment of model fit. *Systematic Biology*, **63**, 334–348.

Coissac E, Riaz T, Puillandre N (2012) Bioinformatic challenges for DNA metabarcoding of plants and animals: Bioinformatic for DNA metabarcoding. *Molecular Ecology*, **21**, 1834–1847.

Coissac E, Hollingsworth PM, Lavergne S, Taberlet P (2016) From barcodes to genomes: extending the concept of DNA barcoding. *Molecular Ecology*, **25**, 1423–1428.

Collins RA, Cruickshank RH (2012) The seven deadly sins of DNA barcoding. *Molecular Ecology Resources*, **13**, 969–975.

De Maio N, Schlötterer C, Kosiol C (2013) Linking Great Apes Genome Evolution across Time Scales Using Polymorphism-Aware Phylogenetic Models. *Molecular Biology and Evolution*, **30**, 2249–2262.

De Queiroz K (2007) Species concepts and species delimitation. *Systematic Biology*, **56**, 879–886.

Dupuis JR, Roe AD, Sperling FAH (2012) Multi-locus species delimitation in closely related animals and fungi: one marker is not enough. *Molecular Ecology*, **21**, 4422–4436.

Galtier N, Nabholz B, Glémin S, Hurst GDD (2009) Mitochondrial DNA as a marker of molecular diversity: a reappraisal. *Molecular Ecology*, **18**, 4541–4550.

Hebert PDN, Cywinska A, Ball SL, deWaard JR (2003) Biological identifications through DNA barcodes. *Proceedings. Biological Sciences / The Royal Society*, **270**, 313–321.

Hebert PD, Stoeckle MY, Zemlak TS, Francis CM (2004) Identification of Birds through DNA Barcodes. *PLoS Biology*, **2**, e312.

Kery M, Royle AJ (2015) *Applied Hierarchical Modeling in Ecology: Analysis of Distribution, Abundance and Species Richness in R and BUGS: Volume 1: Prelude and Static Models*. Academic Press, San Diego.

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.

Kress WJ, García-Robledo C, Uriarte M, Erickson DL (2015) DNA barcodes for ecology, evolution, and conservation. *Trends in Ecology & Evolution*, **30**, 25–35.

Philippe H. (2011b) Une décroissance de la recherche scientifique pour rendre la science durable ? In: Décroissance versus développement durable: débats Pour la suite du monde, Yves-Marie Abraham, Louis Marion et Hervé Philippe, Éditeurs. Écosociété. Pp. 166–186.

Philippe H, Brinkmann H, Lavrov DV *et al.* (2011a) Resolving Difficult Phylogenetic Questions: Why More Sequences Are Not Enough. *PLoS Biology*, **9**, e1000602.

Puillandre N, Lambert A, Brouillet S, Achaz G (2012) ABGD, Automatic Barcode Gap Discovery for primary species delimitation. *Molecular Ecology*, **21**, 1864–1877.

Ratnasingham S, Hebert PDN (2007) bold: The Barcode of Life Data System (http://www.barcodinglife.org). *Molecular Ecology Notes*, **7**, 355–364.

Scheffers BR, Joppa LN, Pimm SL, Laurance WF (2012) What we know and don't know about Earth's missing biodiversity. *Trends in Ecology & Evolution*, **27**, 501–510.

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

Woese CR, Kandler O, Wheelis ML (1990) Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. *Proceedings of the National Academy of Sciences*, **87**, 4576–4579.

Yang Z (1997) How often do wrong models produce better phylogenies? *Molecular Biology and Evolution*, **14**, 105–108.

This work was conceived and written by L.Z. and H.P.

doi: 10.1111/mec.13639