Heredity (2006), 1–2

www.nature.com/hdy

NEWS AND COMMENTARY

Taxonomy

DNA barcodes: recent successes and future prospects

KK Dasmahapatra and J Mallet

Heredity advance online publication, 21 June 2006; doi:10.1038/sj.hdy.6800858

In 'DNA barcoding' a short section of DNA sequence is used to identify species. Neither the idea nor the technology behind DNA barcoding is novel. What is new and controversial is the idea of using just a small portion of a single gene to identify species from a wide taxonomic range, including animals such as birds, fish and insects (Hebert *et al*, 2004b; Ward *et al*, 2005; Hajibabaei *et al*, 2006). This recent usage, and its subsequent successes, has induced criticism and taxonomic debate.

The initial fanfare for DNA barcoding led rapidly to the formation of the Consortium for the Barcode of Life (CBOL, http://barcoding.si.edu), with the objective of obtaining DNA barcodes from all species on the planet. Advances in sequencing technology mean that sequences can now be obtained rapidly and cheaply, so that this barcoding endeavour appears both plausible and worthwhile.

Barcoding has created some controversy in the taxonomy community (Ebach and Holdrege, 2005; Will *et al*, 2005). Traditional taxonomists use multiple morphological traits to delineate species. Today, such traits are increasingly being supplemented with DNA-based information. In contrast, the DNA barcoding identification system is based on what is in essence a single complex character (a portion of one gene, comprising ~650 bp from the first half of the mitochondrial *cytochrome c oxidase subunit I* gene sometimes called *COXI* or *COI*), and barcoding results are therefore seen as being unreliable and prone to errors in identification.

Since the early barcoding papers in 2003 (Hebert *et al*, 2003a, b; Blaxter, 2004), Hebert and co-workers have touted the successes of barcoding in a series of publications. From this single short sequence of the *COI* gene, individuals have been identified down to species level with a success rate ranging from 98 to 100% in North American birds (Hebert *et al*, 2004b), Australian fish (Ward *et al*, 2005) and most recently in tropical Lepidoptera (Hajibabaei *et al*, 2006). As well as correctly identifying known species, a number of probably cryptic species have been discovered within what had previously been thought to be single morphologically based species. This apparent success has fuelled speculation that accurate species identification is now possible by anyone with access to DNA sequencing even if they lack taxonomic expertise.

The methodology used in DNA barcoding has been straightforward. Sequences of the barcoding region are obtained from various individuals. The resulting sequence data are then used to construct a phylogenetic tree using a distance-based 'neighbour-joining' method. In such a tree, similar, putatively related individuals are clustered together. The term 'DNA barcode' seems to imply that each species is characterised by a unique sequence, but there is of course considerable genetic variation within each species as well as between species. However, genetic distances between species are usually greater than those within species, so the phylogenetic tree is characterised by clusters of closely related individuals, and each cluster is assumed to represent a separate species.

Two recent studies convincingly demonstrate the efficacy of DNA barcoding to recover biologically significant groupings or species. Within a single morphologically identified skipper butterfly species, DNA barcoding separates 10 cryptic species (Hebert *et al*, 2004a). These cryptic species differ in larval appearance, food plant or habitat preference. In another recent study, morphologically indistinguishable parasitoid flies (Tachinidae) were shown to be comprised of groups of separate host-specific cryptic species (Smith *et al*, 2006). In these studies, the extraordinary success of the barcoding technique is due to the tight correlation of ecological data with barcoding-based species clusters. This provides compelling evidence that the new species detected by means of DNA barcoding are genuine, rather than merely methodological artefacts.

The utility of barcoding relies on the assumption that genetic variation within a species is much smaller than variation between species. This assumption was valid in the Hebert studies mentioned above and gave 98–100% species identification success rates (Hebert *et al*, 2004a, b; Ward *et al*, 2005; Hajibabaei *et al*, 2006). However, these studies are probably somewhat biased tests of the method. First, intraspecific variation has usually been underestimated, either because only one to two individuals per species were analysed, or because sampling was carried out within a restricted geographic area. Second, interspecific variation has been overestimated because, for most species, sister taxa (their closest relatives) were not included in the analyses as they do not necessarily occur in the areas over which the sampling was carried out. Both of these factors lead to inflated assignment accuracy rates. In contrast to the Hebert laboratory studies, a minimum assignment error of ~17% was reported when comprehensive sampling of both inter- and intraspecific variation within cowries species was carried out (Meyer and Paulay, 2005).

Another problem with using the barcoding region to identify species is that it is located in the mitochondrial genome, rather than the nuclear genome. The nuclear genome contains the majority of genes and is inherited through both parents, while mitochondrial DNA (mtDNA) is only inherited through females. Factors such as interspecific hybridisation and infection by maternally transmitted endosymbionts such as *Wolbachia* are known to cause mitochondrial genes to flow between biological species such that species groupings created using mtDNA can differ from the true species groupings (Hurst and Jiggins, 2005).

To bring greater reliability to the identification of species using short DNA sequences, a move should be made to supplement the mtDNA-based barcode with nuclear barcodes. This would reduce the problem of reliance on a single character and help identify cases where mtDNA behaves differently to the nuclear genome. Most molecular phylogenetic studies routinely make use of multiple nuclear genes, so this is by no means a novel idea. However, there is a snag: most nuclear loci used either evolve too slowly to be useful for distinguishing closely related species, or have intron regions rife with insertions and deletions, and require cloning

to obtain high-quality sequence information from heterozygotes. The challenge is therefore to find 600–1000 bp long nuclear protein coding regions undisrupted by introns, with rates of evolution fast enough to distinguish closely related species.

Despite the obvious drawbacks of using DNA barcoding, the reported success of using the barcoding region in distinguishing species from a range of taxa and to reveal cryptic species is remarkable. However, it is known that species identification based on a single DNA sequence will always produce some erroneous results. Efforts should therefore be made to develop nuclear barcodes to complement the barcoding region currently in use. As the advantages and limitations of barcoding become apparent, it is clear that taxonomic approaches integrating DNA sequencing, morphology and ecological studies will achieve maximum efficiency at species identification.

KK Dasmahapatra and J Mallet are at the Galton Laboratory, Department of Biology, University College London, 4 Stephenson Way, London NW1 2HE, UK.

e-mail: k.dasmahapatra@ucl.ac.uk

Blaxter ML (2004). *Philosophical Trans R Soc London Ser B-Biol Sci* **359**: 669–679.

Ebach MC, Holdrege C (2005). *Nature* **434**: 697.

Hajibabaei M, Janzen DH, Burns JM, Hallwachs W, Hebert PDN (2006). *Proc Natl Acad Sci USA* **103**: 968–971.

Hebert PDN, Cywinska A, Ball SL, DeWaard JR (2003a). *Proc R Soc London Ser B-Biol Sci* **270**: 313–321.

Hebert PDN, Penton EH, Burns JM, Janzen DH, Hallwachs W (2004a). *Proc Natl Acad Sci USA* **101**: 14812–14817.

Hebert PDN, Ratnasingham S, DeWaard JR (2003b). *Proc R Soc London Ser B-Biol Sci* **270**: S96–S99.

Hebert PDN, Stoeckle MY, Zemlak TS, Francis CM (2004b). *PLOS Biol* **2**: 1657–1663.

Hurst GDD, Jiggins FM (2005). *Proc R Soc London Ser B-Biol Sci* **272**: 1525–1534.

Meyer CP, Paulay G (2005). *PLOS Biol* **3**: 2229–2238.

Smith MA, Woodley NE, Janzen DH, Hallwachs W, Hebert PDN (2006). *Proc Natl Acad Sci USA* **103**: 3657–3662.

Ward RD, Zemlak TS, Innes BH, Last PR, Hebert PDN (2005). *Philosoph Trans R Soc B-Biol Sci* **360**: 1847–1857.

Will KW, Mishler BD, Wheeler QD (2005). *Syst Biol* **54**: 844.

Editor's suggested reading

Jolly MT, Jollivet D, Gentil F, Thiébaut E, Viard F (2005). Sharp genetic break between Atlantic and English Channel populations of the polychaete *Pectinaria koreni*, along the North coast of France. *Heredity* **94**: 23–32.

Rohfritsch A, Borsa P (2005). Genetic structure of Indian scad mackerel *Decapterus russelli*: Pleistocene vicariance and secondary contact in the Central Indo-West Pacific Seas. *Heredity* **95**: 315–326.