PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B

Phil. Trans. R. Soc. B (2005) **360**, 1835–1845
doi:10.1098/rstb.2005.1715
Published online 14 September 2005

Wedding biodiversity inventory of a large and complex Lepidoptera fauna with DNA barcoding

Daniel H. Janzen[1,*], Mehrdad Hajibabaei[2], John M. Burns[3], Winnie Hallwachs[1], Ed Remigio[2] and Paul D. N. Hebert[2]

[1]*Department of Biology, University of Pennsylvania, Philadelphia, PA 19104, USA*
[2]*Department of Integrative Biology, University of Guelph, Guelph, Ontario, Canada N1G2W1*
[3]*Department of Entomology, National Museum of Natural History, Smithsonian Institution, Washington, DC 20560-0127, USA*

By facilitating bioliteracy, DNA barcoding has the potential to improve the way the world relates to wild biodiversity. Here we describe the early stages of the use of *cox1* barcoding to supplement and strengthen the taxonomic platform underpinning the inventory of thousands of sympatric species of caterpillars in tropical dry forest, cloud forest and rain forest in northwestern Costa Rica. The results show that barcoding a biologically complex biota unambiguously distinguishes among 97% of more than 1000 species of reared Lepidoptera. Those few species whose barcodes overlap are closely related and not confused with other species. Barcoding also has revealed a substantial number of cryptic species among morphologically defined species, associated sexes, and reinforced identification of species that are difficult to distinguish morphologically. For barcoding to achieve its full potential, (i) ability to rapidly and cheaply barcode older museum specimens is urgent, (ii) museums need to address the opportunity and responsibility for housing large numbers of barcode voucher specimens, (iii) substantial resources need be mustered to support the taxonomic side of the partnership with barcoding, and (iv) hand-held field-friendly barcorder must emerge as a mutualism with the taxasphere and the barcoding initiative, in a manner such that its use generates a resource base for the taxonomic process as well as a tool for the user.

Keywords: Costa Rica; tropical; Area de Conservación Guanacaste; Hesperiidae; Saturniidae; Sphingidae

1. INTRODUCTION

In 1978, D. H. Janzen and W. Hallwachs began the inventory of the entire caterpillar fauna (exclusive of leaf miners) and their parasitoids of Area de Conservación Guanacaste (ACG) in northwestern Costa Rica (Janzen 2000, 2003, 2004*a*; Burns & Janzen 2001; Janzen & Hallwachs 2005; Gauld & Janzen 2004; Hebert *et al.* 2004). Terrestrial ACG is 115 000 ha of dry forest, rain forest, cloud forest, and their intergrades from 0 to 2000 m (http://www.acguanacaste.ac.cr.; Janzen 2000). About 3200 species of caterpillars have now been inventoried (found, reared, photographed, identified, and placed on the project website at http://janzen.sas.upenn.edu.), with approximately 6400 species yet to inventory (as based on a 25-year inventory of adults by Janzen and Hallwachs). This inventory requires a massive ongoing and highly interactive taxonomic platform. It has been provided over five decades by more than 150 members of the taxasphere and their collections, field guides, revisionary papers, and species descriptions, beginning while the senior author was still in high school and visited lowland Mexico to collect butterflies. Interactive revisionary and species-level taxonomy of the inventoried species is the life of the project.

* Author for correspondence (djanzen@sas.upenn.edu).

One contribution of 18 to a Theme Issue 'DNA barcoding of life'.

DNA barcoding for the express purpose of identifying species emerged in 2003 (Hebert *et al.* 2003; www.barcoding.si.edu) as a streamlined, economical, and assembly-line version of the long-established and more general use of DNA sequence information for phylogeny, phylogeography, and population demarcation. We immediately applied it to the taxonomic process underlying the ACG caterpillar inventory. We sought to provide an additional tool for species discovery and identification, as well as to serve as a pilot project for the application of DNA barcoding to complex and species-rich biotas. Byproducts are contributions to the Lepidoptera cytochrome oxidase subunit I (*cox1*) sequence libraries in BoLD and GenBank, stimulation of the eventual emergence of cheap, field-friendly identification barcorders for the world at large, and promotion of the concept of a low-charge-per-individual identification tollbooth that contributes to the financial maintenance of the taxasphere (Janzen 1993, 2004*b*).

2. THE CATERPILLAR INVENTORY PROCESS AND DNA BARCODING

Barcoding fits into the logistics of the ACG caterpillar inventory (methodology at http://janzen.sas.upenn.edu and the Janzen powerpoint deposited at the Consortium for the Barcodes of Life (CBOL) website www.barcoding.si.edu/Presentations.htm) as follows. A free-living caterpillar is found in the forest by one

of the project's 19 resident Costa Rican parataxonomists (Janzen 2004*a*), brought to one of seven rearing barns scattered across the three primary ACG terrestrial ecosystems, and reared through to adult (or parasitoid) in a plastic bag suspended from a clothesline. Its collateral information is maintained as a single event-based record, with the record and the caterpillar assigned a unique alphanumeric voucher code (e.g. 95-SRNP-5116). On its first encounter(s) by the inventory, the caterpillar is photographed. Care of each individual continues until the newly eclosed adult is killed by freezing in a −15 to −20°C non-defrosting freezer. Accumulated adults are removed from the freezer at one- to six-month intervals, their field identifications are corroborated and they are: (i) discarded, (ii) pinned, spread, and oven-dried at 50–60°C, or (iii) placed in 100% ethanol and refrozen or refrigerated. At one-to-six-month intervals, the pinned and dried specimens are hand-carried to the University of Pennsylvania (UP) under a formal export permit from the government of Costa Rica, having been collected under a formal research permit issued by the Ministerio de Recursos Naturales y Energia (MINAE). The latter permit explicitly authorizes the collection of specimens for DNA barcoding. At UP they are sorted for later deposition with participating taxonomists in their respective museums. The legs used for sequencing at the University of Guelph CBOL node are taken from these dry specimens. Likewise, the ethanol-preserved specimens are transported at room-temperature to the University of Pennsylvania and stored again in −20°C freezers or refrigerators, and then donated to specific taxonomic researchers or the Ambrose Monell Collection for Molecular and Microbial Research in the American Museum of Natural History (http://research.amnh.org/amcc) for public scientific use. At the end of each year, the individual databases are pooled from the seven rearing barns, edited and data-checked, pooled with the master database, and posted on the project website. The project currently generates about 35 000 rearing records per year. At the end of 2004, it had logged about seven million caterpillar rearing days, for 264 370 event-based records.

This assembly-line inventory process provides a strong platform for barcoding because:

(i) many conspecific and individually vouchered and databased specimens less than two decades old are museum-available from all ACG ecosystems;
(ii) the inventory voucher specimen is automatically available as the barcode voucher specimen;
(iii) the frozen and then oven-dried specimens have not been field-dried, relaxed at high humidity, and then re-dried when mounted, a treatment that is apparently quite destructive to DNA (occasional specimens are killed with cyanide, but this has had no apparent effect on ease of sequencing) (see Prendini *et al*. 2002);
(iv) each adult moth or butterfly (or parasitic wasp or fly) has three pairs of dry legs and one member of a pair (and yet another in the case of need) can be removed for sequencing;
(v) the specimens are already identified to some level by standard morphology-based or ecology-based taxonomic protocols before entering into the barcoding process;
(vi) when barcoding generates taxonomic questions, the inventory process is modified (as with morphology-based taxonomic processing) to generate more specimens of the taxon in question, albeit with lag times of six months to a year, owing to the intrinsically slow find–rear–eclose process;
(vii) all species being examined are either sympatric within ACG, or, if restricted to different ecosystems, are parapatric at the interdigitations of the ecosystems over distances of a few hundred metres;
(viii) the specimens being compared and identified morphologically are usually in excellent condition, unlike the worn specimens commonly collected as adults; and
(ix) because they are reared, it is often possible to know if a pair of specimens are sibs, and even to use the barcodes of sibs and parents to explore intra-population variation and confirm the accuracy of sequencing.

3. THE FIRST TRIAL

In March 2003, at the first Sloan Foundation-supported conference at the Banbury Centre, we realized that the 'barcoding' initiative (which was to become CBOL at the Smithsonian organizing conference in May 2004) had the potential to be a powerful new tool in the taxonomic toolkit. The ACG inventory sent eight pairs of morphologically similar congeneric skipper butterflies (Hesperiidae) to the Guelph CBOL node. They were found to be easily distinguishable by their *cox1* sequences (termed COI sequences at that time). This prompted Janzen, Hallwachs, and Burns to invite the Guelph node to apply barcoding to an estimated seven undescribed, and morphologically very similar, species detected within ACG *Astraptes fulgerator* (Hesperiidae). Barcoding 484 individuals revealed a total of 10 more or less sympatric species in the complex (Hebert *et al*. 2004).

4. ROUTINE BARCODING IN THE INVENTORY

The clarity of results with *Astraptes fulgerator*, the challenge of applying a new identification tool to the mass of biodiversity information accumulated through nearly three decades of ACG inventory, and the willingness of the Guelph node to barcode tens of thousands of vouchered museum specimens for a few dollars each was irresistible.

(a) *Mechanics of barcoding ACG inventory specimens*

A dry inventory voucher specimen is selected for analysis, and a single leg broken off at its base with forceps. The forceps are tightly wiped with a portion of unsullied Chemwipe tissue between each use. The dry leg is dropped into a new 2 ml Eppendorf tube or into a tube in a 96-tube MATRIX Box (Matrix Technologies,

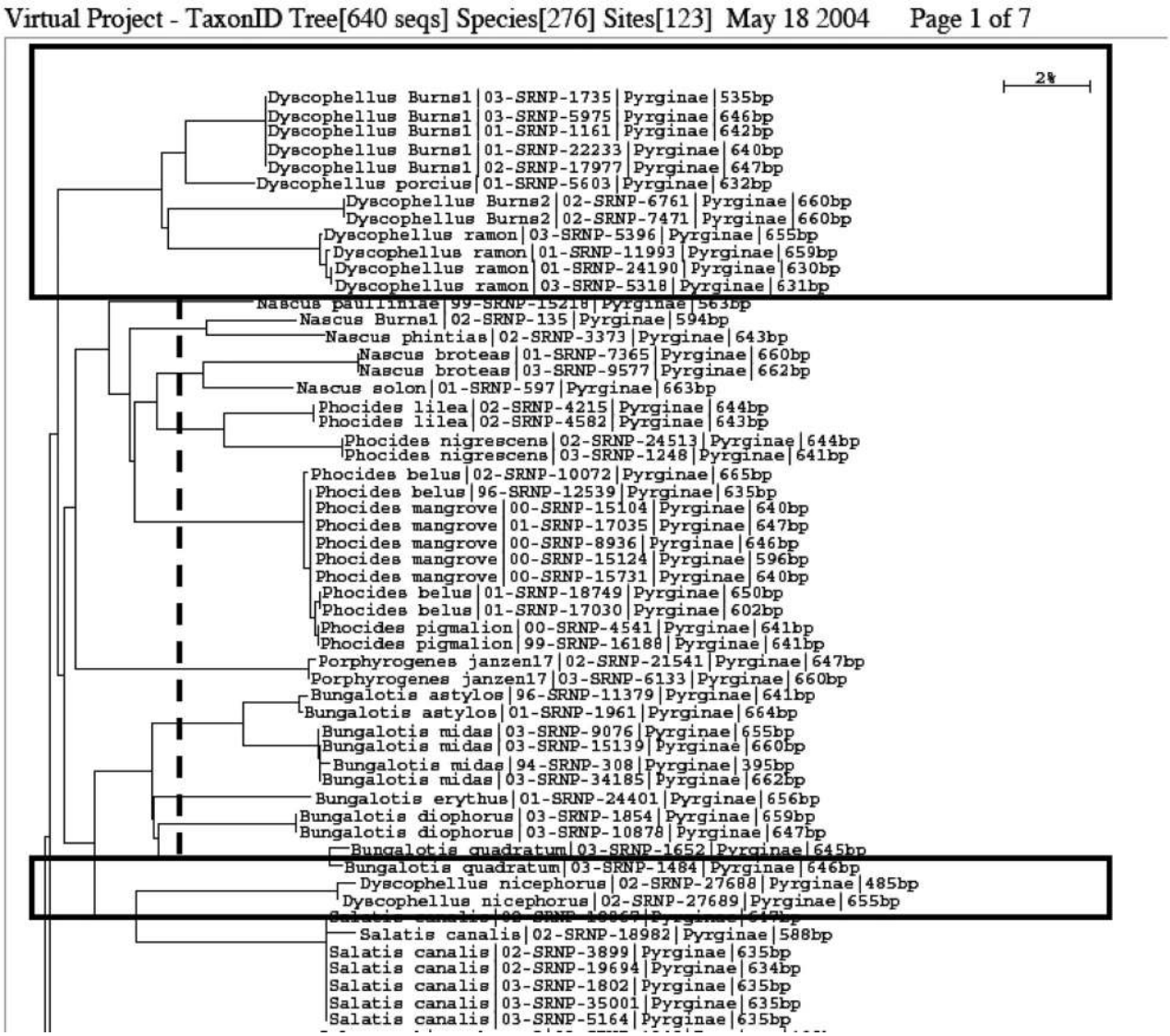

Figure 1. The 18 May 2004 portion of the ACG Hesperiidae NJ *cox1* phenogram containing a grouping of four species of *Dyscophellus* (black frame box) and *Dyscophellus nicephorus* well below that, positioned among *Bungalotis* and *Salatis*.

Hudson, New Hampshire), with a hand-written (India ink on acid-free bond paper) or laser-printed voucher code placed inside the tube, and couriered to the Guelph node. The museum specimen is flagged with a yellow 'legs away for DNA' pin tag, as is the voucher database record. The voucher specimen's collateral information is uploaded from the inventory database to an Excel form prepared by the Guelph node, and accompanied by two images (upperside and underside), all of which are placed in the specimen's record in the project databases at Barcode of Life Database (BoLD) at www.barcodinglife.com.

At the Guelph node, DNA is extracted from each leg and *cox1* ('COI' in previous literature) is PCR amplified and sequenced. The *cox1* sequence is placed in BoLD for processing, and later submission to GenBank, along with its collateral information. Residual DNA extracts are preserved in −80°C freezers. Specimens that do not sequence well are variously re-sequenced and otherwise processed, depending on the question being asked (see Hajibabaei *et al.* 2005). The ACG inventory subsequently obtains the placement of this specimen relative to others by constructing a Neighbor Joining (NJ) phenogram (a 'species identification phenogram') by using the BoLD website (see examples below and Hebert *et al.* 2003; Hajibabaei *et al.* 2005). The NJ phenogram can have bootstrap values placed on it if relevant, and the specimen's position can be labelled with voucher code, name, geographic location, higher taxon, and/or sequence length as the project wishes. Different subsets of specimens may be differently coloured at the command of the user. The user can also download individual sequence data and collaterals. At the current evolving and developing process at the Guelph node, this entire process costs the ACG inventory $2.50/specimen once it arrives in Guelph. This extremely low price is, however, achieved by subsidy from other grants, most notably from the Gordon and Betty Moore Foundation, the Canadian government, and the University of Guelph.

(b) *Typical results*

The 2640 specimens of the ACG-reared Hesperiidae of about 350 species barcoded to date produce a manageable NJ phenogram (Electronic Appendix)

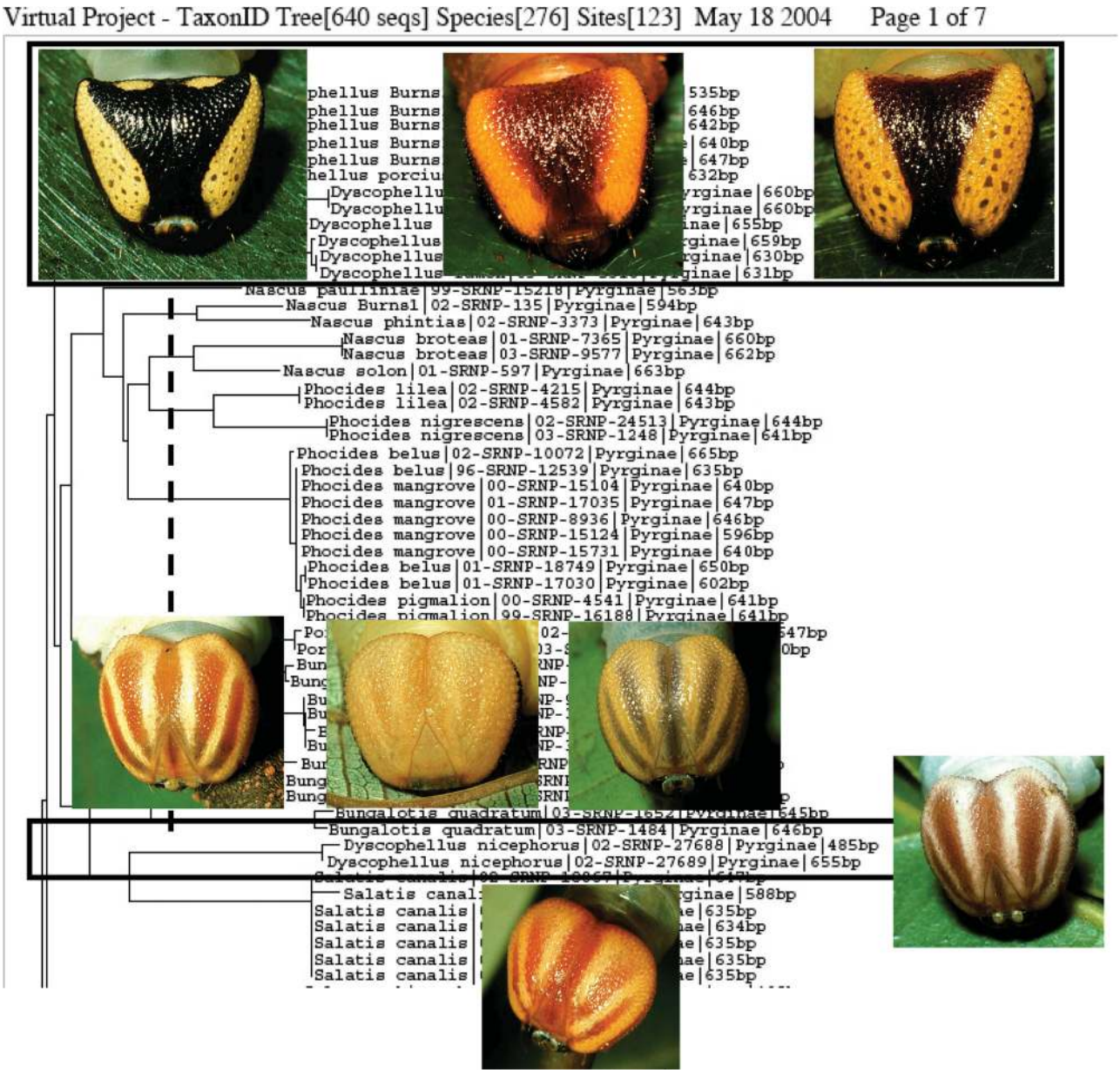

Figure 2. As in figure 1, but with the colour patterns of last instar caterpillar heads superimposed on the phenogram. *Dyscophellus nicephorus* is offset on the lower right.

that illustrates many practical aspects of barcoding in this inventory. Figure 1 highlights the portion of this NJ phenogram containing the sequences from 12 specimens of four species of sympatric rain forest *Dyscophellus*, two of which have similar facies but are readily distinguishable by their genitalia. Two are undescribed and, therefore, bear interim names. Three of the four can be easily distinguished by their caterpillar-and-food-plant combinations. A similar level of separation between congeneric species in the NJ phenogram occurs with about 97% of the 1000-plus morphologically defined ACG species sequenced to date in Hesperiidae, Saturniidae, Sphingidae, Nymphalidae, and Arctiidae. As the sample size for each species increases, the clusters in the NJ phenogram retain their species-level discreteness. The placement of a sequence from an unidentified ACG specimen into one of these clusters means that it is very likely to be that species, unless it is a previously unknown species that is among the 3% of confusables (see below).

(c) *Phylogenetic signals?*

While barcoding does not aim to build phylogenetic trees, it is obvious that morphology-based congenerics are often the nearest neighbours in the NJ phenogram. When they are not, it is a signal that the morphological placement may be profitably re-examined. With respect to the example of four species of *Dyscophellus* given earlier, a fifth sympatric species, *Dyscophellus nicephorus*, appears well removed in the NJ phenogram, among the array of *Bungalotis* and *Salatis* (figure 1). Despite the similarity of adult facies of *Dyscophellus nicephorus* to the other four *Dyscophellus*, some members of the inventory staff have long suspected that it was misplaced because its caterpillar has the same colour patterns as do *Bungalotis* and *Salatis*, rather than the distinctive colour pattern of the other

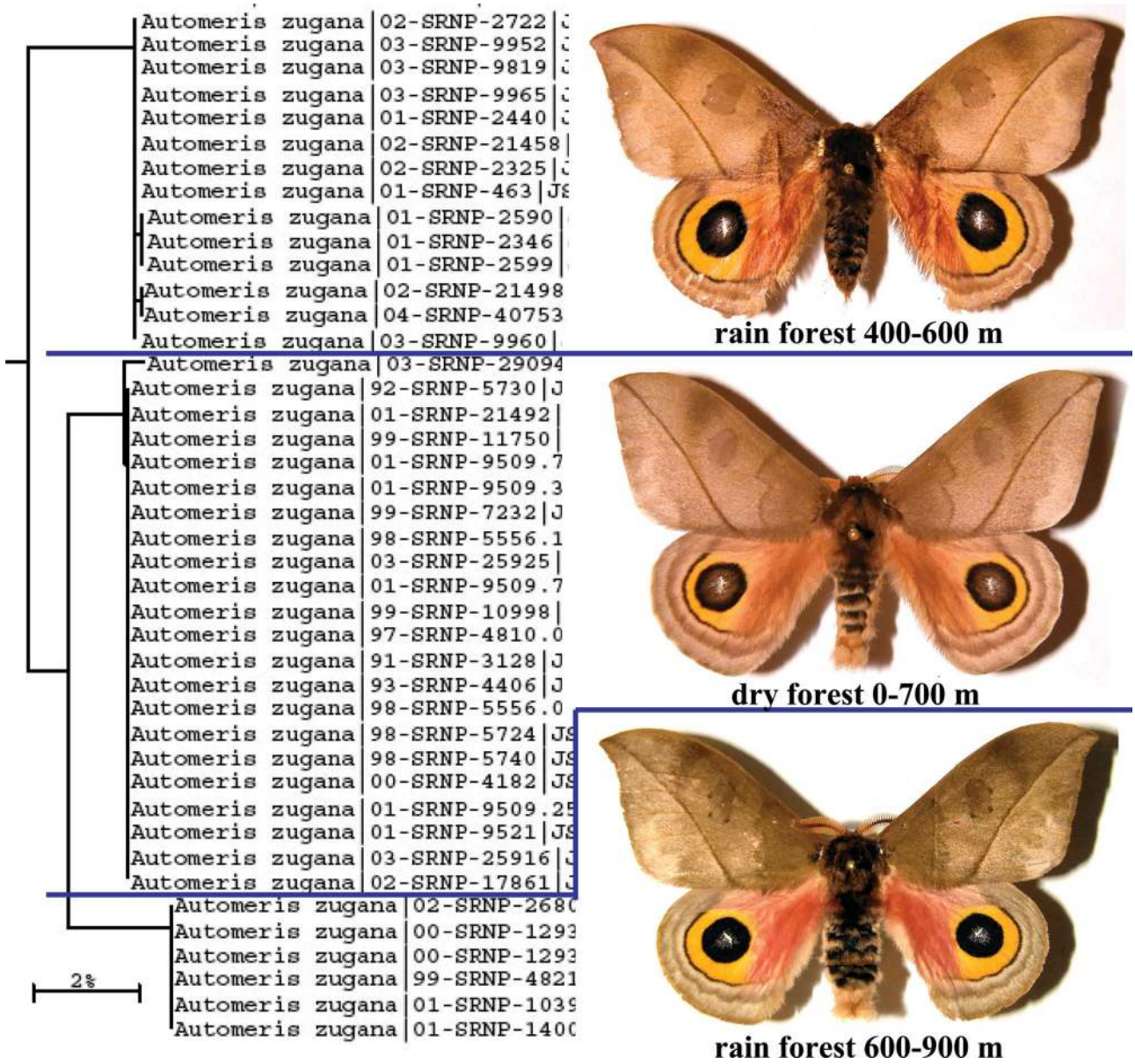

Figure 3. The 16 March 2004 portion of the ACG Saturniidae *cox1* NJ phenogram containing the three cryptic sympatric species within what has been called *Automeris zugana* and revealed by barcoding. Each specimen is a male of a different species.

four *Dyscophellus* (figure 2). Similarly, the sixth ACG congeneric, *Dyscophellus phraxanor*, has an adult female and a caterpillar that matches well with the four similar *Dyscophellus*, but a very different male; this species also positions far from all of the others in the NJ phenogram (Electronic Appendix). Barcoding unambiguously distinguishes among the six species of ACG *Dyscophellus*, does not confuse them with any other Lepidoptera examined, and suggests that some of their generic placements should be re-examined.

(d) *Morphological species indistinguishable by barcoding*

About 3% of the 1000-plus morphological species of ACG Lepidoptera that have been barcoded to date cannot be distinguished from a close relative by their barcodes. An example is three species of *Phocides*. They are distinguishable by wing patterns, genitalia, and caterpillar food plants; but their barcode positions intermingle in the NJ phenogram (Electronic Appendix). However, they neither intermingle with the other three species of ACG *Phocides*, nor with the six other species of look-alike ACG Hesperiidae in two subfamilies and four genera. Other cases of a lack of barcode resolution of ACG hesperiids include *Saliana fusta* and *Saliana triangularis*, and *Cobalus virbius* and *Cobalus fidicula*; two sphingid examples are *Cautethia spuria* and *Cautethia yucatana*, and *Manduca lanuginosa* and *Manduca barnesii* (note added in proof: *Manduca confusion* now appears to be due to sample contamination); there are no saturniid examples (Electronic Appendix).

(e) *Morphological species with very similar barcodes*

The morphological species barcoded to date offer a few cases where morphologically similar species possess distinct but very similar clusters (differing by less than 1%) in the NJ phenogram. A dramatic example is offered by *Polyctor cleta* (ACG dry forest) and *Polyctor polyctor* (ACG rain forest). These two medium-sized hesperiids are extremely similar but distinguishable by

facies and genitalia. They differ by just four base pairs in the 648 base pairs region. The single potential sphingid example (Electronic Appendix) is within *Xylophanes crotonis*. Even here, it is unclear as to whether the two clusters within this species in ACG should be viewed as two morphologically identical sympatric species or merely a pair of equally common *cox1* polymorphisms within a single species. Parenthetically, neither of the two clusters represents the newly described other *Xylophanes crotonis* look-alike, *Xylophanes letiranti* (Vaglia & Haxaire 2003), which occurs near but not in ACG. There are no cases of very similar barcodes among morphologically defined ACG-reared Saturniidae.

(f) *Dissolution of one morphological species into several*

Apart from the exceptionally species-rich case of *Astraptes fulgerator* becoming 10 ACG species (Hebert *et al*. 2004), the barcoding of reared ACG Hesperiidae, Saturniidae, and Sphingidae contains significant numbers of examples of an apparent morphological species becoming two or more clusters of adjacent barcodes in the NJ phenogram. On close inspection of their food plants, behaviour, ecosystem or elevation occupied, and/or adult morphology, many—but not all—of these clusters are being found to represent distinct biological entities in ACG.

An example is *Automeris zugana*—a medium-sized, widespread and very well-known saturniid moth (Costa Rica to Ecuador, Lemaire 2002). The first three specimens barcoded, chosen deliberately to span the dry forest and rain forest sides of the ACG, displayed a 2–4% difference in their sequences. While these are substantially smaller differences than those among most morphologically defined species of ACG saturniids (Electronic Appendix), they were large enough to suggest hidden complexity. When 10 *A. zugana* were barcoded, three distinct clusters of sequences emerged. When 42 specimens, chosen to cover the ACG ecosystems, were sequenced, the clusters unambiguously remained (figure 3). The three clusters correlate with subtle differences in adult body weight, facies, genitalia, and ecosystems (the caterpillars are indistinguishable, as are their food plant preferences). The morphological differences had been viewed as intraspecific variation at the time that wild-caught adults were examined by the inventory and by the late Claude Lemaire in the 1980s, though we suspect that Lemaire did not examine the genitalia of more than a few ACG specimens, which happened to be of just one species. One barcode cluster occupies the ACG dry forest, and two occupy the adjoining rain forest—one at 400–600 m elevation and the other at 600–900 m. Ongoing taxonomic efforts will probably link one of these three species to the type specimen of *Automeris zugana* and describe the other two as new. Once described, these three species would fall in the category above of species that differ only slightly in their barcodes but are readily distinguishable by their barcodes.

While examples like that of *A. zugana* are not unusual among the hard-to-catch and often-low-density Hesperiidae (more than half of the Hesperiidae reared by the caterpillar inventory have never been seen or collected as adults in ACG), they were less expected among Saturniidae and Sphingidae, so loved by collectors and so easily collected with light traps. As mentioned earlier, *Xylophanes crotonis* might turn out to be one of these cases. Other cases still being explored are potential cryptic species within *Xylophanes porcus*, *Xylophanes libya*, *Manduca sexta*, and *Pachylia ficus*—four seemingly well-known and widespread morphological species. Several other well-known ACG saturniids are experiencing the same fate as described for *Automeris zugana*. There are two unexplored barcode clusters within *Gamelia musta*, *Automeris tridens*, *Automeris postalbida*, and *Hylesia dalina* (Electronic Appendix). All four of these hemileucine saturniids are highly polyphagous as caterpillars (Janzen 2003), and the barcode clusters are parapatric by ecosystem and/or elevation. The most startling of all is the well-known *Eacles imperialis*, which ranges from southeastern Canada to Argentina. The ACG *Eacles imperialis* has two distinct barcode clusters showing an 8% sequence divergence. One cluster occurs in rain forest and the other in the parapatric dry forest. Strikingly, the dry forest cluster only differs by 5% from its morphological conspecific in Great Smoky Mountains National Park, Tennessee, USA, several thousand kilometres to the north (sequences from BoLD). *Eacles imperialis* do not migrate.

If the adult Hesperiidae, Sphingidae, and Saturniidae of ACG had not been so thoroughly studied morphologically during the past 100-plus years, there would be many more cases where 'one' slightly variable morphological species dissolves into several when it is barcoded.

(g) *Association of sexes*

Associating sexes of wild-caught or reared polyphagous species-rich Lepidoptera can be difficult. Scott Miller and colleagues have already found barcoding to be extremely useful in associating sexes of their reared Tortricidae and Lymantriidae in their extensive caterpillar inventory in Papua New Guinea (www.nmnh.si.edu/new_guinea). In the ACG inventory, the caterpillars of two distinctive 'species' of *Saliana* (Hesperiinae) were found at low density, the adults of one being given an interim name and the other tentatively identified as *Saliana severus*. This is an exceptionally dark species of *Saliana*. Barcoding then showed that these two morphological entities had identical *cox1* barcodes. Querying back to the morphological taxonomy, it was noticed that *both* sexes of *Saliana severus* have dark undersides, and that the interim white-undersided *Saliana* were all females, while the ACG *Saliana severus* were all males. This iterative feedback led to the conclusion that the inventory is not rearing *Saliana severus* but yet some other species of *Saliana* with strong sexual dimorphism.

(h) *Massive interspecific discrimination*

BoLD now contains thousands of vouchered and species-level identified *cox1* sequences from 1000-plus species from the ACG inventory, and has accumulated similar records from another 2000 Lepidopteran species from other parts of the world. This leads to

the obvious experiment of comparing all the ACG specimens in one huge NJ phenogram. We did, and there is no overlap of any species other than those already found with a within-ACG family-level NJ phenogram. Next, we combined all BoLD Sphingidae sequences from Africa ($n=26$ for 11 species), Papua New Guinea ($n=75$ for 28 species), North America ($n=136$ for 32 species), and ACG ($n=614$ for 95 species). Again, there is no overlap of the 166 morphological species clusters in the NJ phenogram other than the less than 3% already recognized as confusable within a geographic region.

5. CAVEATS AND PROBLEMS

Combining barcoding with the more classical taxonomic process for the inventory in ACG, and serving as a pilot project for barcode library construction, barcorder emergence, and tollbooth development, is a work in progress. Some barriers to progress have emerged.

(a) *Sample size per species*

It is now commonplace to use mtDNA sequence data to resolve phylogeography of species (e.g. Wuster *et al.* 2005). However, there has been a strong tendency in barcoding to treat a few sequences as if they were the 'type' for a place, potentially missing cryptic species and cases of overlap in the NJ phenogram. This approach was due to an initial desire to maximize species coverage at a time when sequencing costs were still high and analytical protocols were under development. The barcoding done to date with morphologically defined species suggests that if only two to five specimens are barcoded, cases of interspecific overlaps will be recognized; but a significant number of cryptic species that differ by only a few per cent will be missed. While further barcoding is needed to refine this estimate, at least 10 specimens per species should be used from what seems to be one site—assuming that the specimens can be chosen so as to avoid sibling individuals. Samples of this size should expose clues to most cases of sympatric cryptic species that have species-level barcode differences. If such a sample reveals more than one cluster in the NJ phenogram, additional specimens should be barcoded to explore for cryptic species.

(b) *Barcoding a morphologically unknown biota*

The specimens barcoded in the caterpillar inventory are all sorted to morphospecies (often backed by a species-level name) before the specimens are chosen to be barcoded. This minimizes the number of individuals necessary to barcode in order to know how many clusters there are in the NJ phenogram for any given per cent difference used to define a cluster. It also assists in knowing how to treat singletons that deviate slightly from other members of a cluster but do not form or join a cluster. Are they singletons of a rare species or simply deviant individuals? If barcoding simply examines a pool of individuals collected in a Malaise or light trap, a much larger number of individuals would need to be barcoded to reveal all the clusters in the sample. Furthermore, a small fraction of the individuals would remain in taxonomic limbo because it would not be clear if they were the result of intraspecific variation or rare individuals of another species. This is just as it is with morphological sorting of a large sample of unknowns. However, combining barcoding with morphological sorting will give both a more accurate and a more economic result.

(c) *Cross-geography barcoding*

The ACG inventory and its barcoding is, and will continue to be, a deep sample of a place where any sample point is within flight distance of most other sample points. It does not reveal the extent of intraspecific variation in barcodes that will emerge as widespread species are barcoded across their neotropical ranges (e.g. Dick *et al.* 2004). However, this work is well underway for moths and butterflies in the eastern half of North America. Early results suggest that between-site intra-specific variation in barcodes will not be a confounding problem in their use for species identification, except in the very small percentage of species whose barcodes overlap.

(d) *Developing barcoding versus using barcoding*

In the CBOL barcoding initiative (http://www.barcoding.si.edu/), as during the emergence of any new technology, those embedded in the initiative are caught in a tension between full-blast development of barcoding (how to sequence accurately and cheaply, build the sequence libraries, build the barcorder, build and operate the toll booth), and using the new information to solve questions and drive initiatives in other agendas. When do we stop using barcoding to better the caterpillar inventory and be a pilot project, when do we put full time into building the sequence library—with museum and fresh-caught specimens—to barcode the Lepidoptera of the world? The question hinges on availability of funds/technology for each route, on the existence of fellow travellers, and on the personal curiosity yield from each of the two routes. Janzen and Hallwachs are caught up in the mosaic of agendas cocooning the survival of ACG into perpetuity and its pilot project role in biodiversity survival through non-damaging development (Janzen 2000). Burns and the remainder of the taxasphere are caught up in the business of the taxasphere. The CBOL node at Guelph and its occupants (e.g. Hajibabaei *et al.* 2005) are certainly on the barcoding route, but even they will be distracted from the straight and narrow of developing barcoding as a process and into the application of that process to the real world, if for no other reason than to keep the funding flowing.

This study is a microcosm of this problem. Each time a new array of ACG specimens is barcoded, new taxonomic and biodiversity puzzles are revealed. Each begs for taxonomic, ecological, methodological, and publishing energy for its resolution. Barcoding reveals such puzzles at a far higher rate than they can be treated by the human and financial resources available. This means that, for the sake of barcoding, they are left behind. An example is the publication of the barcoding confirmation and exposure of 10 species in ACG *Astraptes fulgerator* (Hebert *et al.* 2004), before the

species have been described and their (known to the caterpillar inventory) natural histories recorded in print or website. All signals are that full-scale barcoding will reveal innumerable questions, as did the microscope and scanning electron microscope. The ACG caterpillar inventory itself has already been heavily exposed to this conundrum, and has resolved it by using a website database to record the basic specimen-level information rather than an interminable series of short publications. It also refuses to be diverted from the goal of total inventory. As frequent observers of taxonomists identifying museum specimens, Janzen and Hallwachs have particularly noticed the positive feedback when the barcoding process is applied to previously studied specimens and identifications. It is a real joy to watch the outcome of providing a top-flight taxonomist with a new tool to address long-standing taxonomic tangles and uncertainties. But that very positive outcome is also highly seductive away from continued development of barcoding as a tool and method.

(e) *Variants*

Anticipation of problems with barcoding leads immediately to concern about hybrids. However, a hybrid should simply cluster with its mother, grandmother, sisters, etc. in a *cox1*-based NJ phenogram. This is no worse than occurs with a morphological search for hybrids. More puzzling are the moderately frequent cases in the ACG inventory where a single individual differs from the remainder of a large sample cluster by two to eight base pairs, but lacks any morphological or natural history reason to be suspected as an individual of a cryptic species, and does not join any other cluster as the sample size is increased (evident examples in Electronic Appendix). These cases may simply be 'deep intraspecific variants' similar to those encountered regularly in morphological and behavioural explorations, but they do beg for a more scientific explanation.

(f) *Laboratory errors*

The processing chain from a caterpillar to a sequence in GenBank (http://www.barcoding.si.edu/CBOLDatabasesGenBank.htm), with its collateral information attached, offers a wealth of opportunities for human and machine errors to creep in. Many of these opportunities, as well as the specific errors themselves, are polished out of the system as discovered on a case-by-case basis. However, there is one general problem that needs immediate attention. It is essential that the internet connectivity among the various data and specimen deposits become so seamless that an error encountered in a data point or its collateral at one place in the chain can be corrected, and then that correction is automatically transmitted through the network to the other places where the uncorrected data remain. To emphasize this need is not a great intellectual advance, but rather a plea for rapid resolution. As we attain consensus that all DNA sequences, for example, should be vouchered with specimens, collateral information, and images, we desperately need to avoid each node in the chain being a static depository of errors that were corrected in one place but cannot be corrected elsewhere without enormous investment of painstaking human and case-by-case intervention at other nodes. A specimen of *Astraptes* TRIGO is simultaneously a sequence and its collaterals in GenBank and BoLD, on a pin in the National Museum of Natural History, Smithsonian Institution, and an event record in the caterpillar inventory. When it finally gets its scientific name, it is imperative that with a push of a button *Astraptes* TRIGO disappears, other than in audit trails, and is everywhere replaced by its newly assigned scientific name. Equally, when it is found that a base pair read was wrong in its deposited sequence, whether by a person or an application, the subsequent corrections throughout the network need to occur not by event-specific emails, but by hot linkages, with all that implies.

(g) *Old specimens*

The only practical way to rapidly build thorough and cross-geography global barcode sequence libraries of millions of species is by barcoding representative specimens in the world's museums. The barcode initiative is not going to recollect the world to build its sequence library. CBOL has done a magnificent job of getting the world's museums politically on board, but there are two major impediments. First, the funds, personnel, and energy are not yet available for the massive taxonomic and physical curatorial process that is required. Worse, the present process will constantly be caught in the dilemma described above whereby the participatory taxonomist is forced to choose between pursuing the multiple taxonomic puzzles and answers revealed by barcoding 'the collection', and sustaining the humdrum of minimal curating for barcoding. This begs for funding for a new kind of curator who largely carries forward the barcoding process while the taxonomist energy is applied surgically to select questions. Even these will quickly exhaust the current taxonomic human resource. The situation absolutely demands an absolute increase in the taxonomist guild if barcoding is to function. Just the questions generated by barcoding the ACG caterpillar inventory can easily absorb the full taxonomic capacity of several major museums for the caterpillar family in question, and ACG contains no more than 3% of the world's Lepidoptera biodiversity.

Second, while sequences can be obtained from a given old specimen with much work and time (and money), we are still far from the fast cheap sequencing that can be done with freshly collected material (however, see Hajibabaei *et al*. 2005). This deficit is a composite of two problems. On the one hand, because fresh material—such as that reared by the ACG inventory—is so easy to sequence (and often is fully databased and vouchered from the beginning), it seduces the barcoding initiative away from the essential ability to analyse the old but much more biodiverse material sitting in museum cabinets and representing a huge geographic coverage and centuries of effort. On the other, if a taxonomist does devote extra curatorial and databasing effort to organize a museum's holdings for barcoding, but only a small fraction of samples successfully sequence, the negative psychological impact is huge. Equally bad is the damage and cost of having to sample a very large number of specimens with

a hope that a few of them will successfully sequence. Incidentally, one of the huge advantages of being able to sequence for barcoding right at the museum cabinet would be that *if* it fails, it is known right then and a second sample or specimen can be tried from the same series. This is much better than having to relocate the failed specimen or species months later among its millions of compatriots. Likewise, onsite sequencing will reveal variation and cryptic species at the time they are being curated, allowing sequence sample size to be increased at that moment.

(h) *Museums and databases as voucher depositories*

It is imperative that barcode sequences be vouchered by specimens, irrespective of whether the specimen has been identified. And as the vouchers become identified, the value of the barcode sequence increases greatly (e.g. De Ley *et al.* 2005 in this Theme Issue). When the specimen is already in a museum for other purposes, making a barcode voucher of it may mean relatively little change in its cost of permanent maintenance. However, the massive barcoding of new inventory specimens, just as the inventory itself, can easily swamp the holding capacity of our museums. Worse, it can do it with huge series that have large barcoding significance for geographic variation, etc., but are far beyond the traditional reasons and amounts of space allocated in museums to long series of conspecifics. The barcode vouchers from the ACG caterpillar inventory have the potential to consume a substantial amount of the new drawer space in the new expansion and reorganization of the Lepidoptera collection of the National Museum of Natural History at the Smithsonian Institution. Were ACG to take on a total Lepidoptera barcode venture, it would require another depository solution. Equally, the Lepidoptera collection at the Instituto Nacional de Biodiversidad (INBio), Costa Rica's National Biodiversity Institute, is filled to capacity. To thoroughly barcode the Lepidoptera of Costa Rica would require a doubling of space at INBio just to hold the vouchers.

The problem is compounded when barcode vouchers are viewed as stored permanently, which means a huge archival cost with no more scientific return than confirmation capacity for a sequence. It seems clear that true vouchering both for barcode libraries, and for research barcoding once a basic library is established, will require the creation of depositories for that purpose rather than simply squeezing more specimens into currently overcrowded museum facilities. A related question is whether a museum is willing to let a taxonomic specimen be moved into the category of barcode sequence voucher, thereby limiting many of the traditional uses for a specimen. The ACG inventory specimens are gladly given to museum repositories as barcoding and taxonomic vouchers, but trading them, resampling them, displaying them and generally caring for them as individually coded vouchers is a major responsibility not to be entered into lightly for the tens of millions of specimens that true global barcoding implies.

6. CONCLUSIONS

DNA barcoding, as being practised on the ACG caterpillar inventory, is about cheap, mass, and fast sequencing to initially discover and confirm biological species, build reference sequence libraries for the species treated, and eventually use the reference library to aid species-level identifications. The cheaper and quicker it is, the easier it will be to explore the complexity of barcode patterns—and the biology they signal—in time and space.

In a world lacking the taxasphere, the single largest problem with barcoding is the inability to connect the cluster in the NJ phenogram to what is already known about that species by humanity. Barcode reference libraries based on, and connected to, what we already know are essential. But what of the millions of species that can be recognized only through a barcode either because they are very similar morphologically, or because they simply have not been studied enough to know their non-barcode diagnostic traits? These species will simply have to exist in some higher taxonomic rank until they are studied as biological entities, and/or until there truly is a pocket barcorder that is used just as are today the camera, hand lens, dissecting microscope, binoculars, notebook, paper field guide, memory, etc. A barcorder is a DNA microscope with a memory. Given the high potential for the barcorder to store every sequence read, along with the collateral of the moment, there is truly huge potential identificatory power and ability to connect to what is locally to globally known. Historically, it should be recognized that a barcorder is far from being the first effort for an automated and computer-based species identification tool. Classical keys up through complex web-based interactive keys, though based on a taxasphere-derived terminology, are themselves a kind of NJ phenogram. A recent example is DAISY, an automated identification concept and tool based on image data rather than the DNA barcode sequence (Gauld *et al.* 2000).

Ongoing integrations of barcoding with field and museum biodiversity studies make clear the need for five 'libraries'—the 'literature', morphology, natural history (food plants, microgeography, phylogeny, etc), taxonomy *per se*, and DNA barcode sequences—and merge them iteratively to approach reality and biolit-eracy. Each of these five libraries is imperfect and variously developed, but when they are merged, they jointly achieve about as good a focus on the biology of a place or taxon as can be obtained.

Apart from the general scientific and public desirability to be able to better, faster, and more cheaply identify organisms for a host of agendas, is there an additional reason to hasten to a realized barcorder and accompanying information? Those of us who would like to see a serious part of today's surviving biodiversity still with us centuries from now are in a severe race against the multiplex of forces polishing today's remnants of that biodiversity off the earth. While it is certainly not the solution to end all concerns, a cheap public back-pocket barcorder does have the potential to allow any and all to know what an organism is at the moment that it matters. This essentially allows anyone to 'read' biodiversity. As with most literacy, it is

only at certain key times that it matters. But if people can be bioliterate at those key times, humanity's relationship to wild biodiversity has a high potential of changing for the better. Yes, there will be abusers, just as there are abusers of literacy, but overall, becoming literate has had a highly civilizing impact on humanity. And from the quite selfish viewpoint of the practising biologist, it greatly increases the motivation to collate and organize what we know if the world can get to that information, even if only on the web, at the moment when the actor in the play is biting, stinging, pollinating, munching, or displaying.

However, it is no secret to the world of users and protectors of wild biodiversity that their politico-legislative framework is built on a taxonomic structure that variously defines species (and their subunits), and usually does it morphologically. Barcoding is going to reveal and reinforce a lot of cryptic diversity, and add fuel to the argument of whether we are using or protecting a morphologically defined or a phylogenetically defined biological entity (e.g. Agapow *et al*. 2004; Debrunye 2005; Simmons *et al*. 2005). Like any broadly applicable technology, it will be used for bad and good; the barcoding initiative will need to be prepared for that. It was correctly anticipated in 2003 that national permission to barcode thousands of species in the ACG would require years of Costa Rican political debate and permission, legislative interpretation, and explicit enlightenment of social leaders.

In the search for rational support for DNA barcoding—as if any is needed other than its obvious pragmatic usefulness—it has been expressed that a major 'problem' with taxonomy is that there are few taxonomists and that one cannot manage more than a few thousand species in his or her head. Both statements are false. There are many taxonomists, but very few jobs for them. Worse, many of these jobs require that they spend substantial time and mental energy on other tasks than taxonomy. We do not need to train more taxonomists so much as we need to hire more of them—the taxasphere combined with individuals who really enjoy doing taxonomy will provide the human resource if there is employment available. Second, we know many taxonomists who handle accurately tens of thousands of names in the combinations of their heads, databases, collections, and literature. Mental capacity is not the problem. The problem is that there is not one of them standing by your left elbow when you need to identify something. And there never will be, no matter how appreciative society becomes of wild biodiversity. A cheap thorough pocket barcorder, and all its supporting information, technology, and linkages, is the only way that the grand bulk of humanity will ever become bioliterate, at least to the degree where living things are generally viewed as more than more biomass to convert or trash.

We close with a reiteration of four speed bumps for the CBOL initiative:

(i) Cheap and fast barcode sequencing of old specimens needs to be developed quickly.
(ii) Museums, the taxasphere, and the user community need to decide if they are willing to take on the permanent housing/storage/curation of the massive numbers of voucher specimens that will be generated by building true global barcoding sequence libraries.
(iii) Funding is essential for the interactive classical taxonomy and curation to provide and name the specimens that will be used to build the DNA sequence libraries. Finding people is not nearly as large a problem as is finding the salary and operational support for the people that already have a strong interest in being participants. We need to HIRE more taxonomists. They will train and mentor each other and themselves. And every time a leg, feather, or leaf chip goes into a barcorder, a tollbooth has to move a penny into the funding for the taxasphere.
(iv) Someone has to take up the conversation with the commercial sectors such that while the barcorder is being built, the emerging technology is in a conversation with the sequence libraries and the tollbooth. A marvellous cell phone is of no use if, when you call the number, no one answers, and, when they do, they have no information.

This study has been supported by grants to DHJ and WH from the Wege Foundation, US National Science Foundation (DEB 0072730), Guanacaste Dry Forest Conservation Fund, and Area de Conservacion Guanacaste. JMB was supported by the Smithsonian Institution and the National Museum of Natural History Small Grants Program. PDNH thanks the Gordon and Betty Moore Foundation, the Canada Foundation for Innovation, the Ontario Innovation Trust, the Canada Research Chairs Program, and NSERC for the critical support that they provided to develop and operate a DNA barcoding facility. We also thank the ACG parataxonomists (Janzen 2004*a*) for assembling specimens, as well as Tanya Dapkey and Stephanie Kirk for acquiring barcode sequences, and Donald J. Harvey for genitalic dissections and Young T. Sohn for drawings. We deeply appreciate the contributions made by Rob Dooh and Sujeevan Ratnasingham to data organization and assembly. We thank M. Stoecklem, S. Miller, V. Savolainen for constructive commentary on the manuscript, and are grateful to CBOL for facilitating the DNA barcoding initiative.

REFERENCES

Agapow, P. M., Bininda-Emonds, O. R. P., Crandall, K. A., Gittleman, J. L., Mace, G. M., Marshall, J. C. & Purvis, A. 2004 The impact of species concept on biodiversity studies. *Q. Rev. Biol.* **79**, 161–179. (doi:10.1086/383542.)

Burns, J. M. & Janzen, D. H. 2001 Biodiversity of pyrrhopygine skipper butterflies (Hesperiidae) in the Area de Conservación Guanacaste, Costa Rica. *J. Lep. Soc.* **55**, 15–43.

De Ley, P. *et al*. 2005 An integrated approach to fast and informative morphological vouchering of nematodes for applications in molecular barcoding. *Phil. Trans. R. Soc. B* **360**, 1945–1958. (doi:10.1098/rstb.2005.1726.)

Debruyne, R. 2005 A case study of apparent conflict between molecular phylogenies: the interrelationships of African elephants. *Cladistics* **21**, 31–50.

Dick, C. W., Roubik, D. W., Gruber, K. F. & Bermingham, E. 2004 Long-distance gene flow and cross-Andean dispersal of lowland rainforest bees (Apidae: Euglossini) revealed by comparative mitochondrial DNA phylogeography. *Mol. Ecol.* **13**, 3775–3785. (doi:10.1111/j.1365-294X.2004.02374.x.)

Gauld, I. D. & Janzen, D. H. 2004 The systematics and biology of the Costa Rican species of parasitic wasps in the thyreodon genus-group (Hymenoptera: Ichneumonidae). *Zool. J. Linn. Soc.* **141**, 297–351. (doi:10.1111/j.1096-3642.2004.00116.x.)

Gauld, I. D., O'Neill, M. A. & Gaston, K. J. 2000 In *Driving Miss Daisy: the performance of an automated insect identification system* (ed. A. D. Austin & M. Dowton) *Hymenoptera: evolution, biodiversity and biological control*, pp. 303–312. Canberra: CSIRO.

Hajibabaei, M., de Waard, J. R., Ivanova, N. V., Ratnasingham, S., Dooh, R. T., Kirk, S. L., Mackie, P. M. & Hebert, P. D. N. 2005 Critical factors for assembling a high volume of DNA barcodes. *Phil. Trans. R. Soc. B* **360**, 1959–1967. (doi:10.1098/rstb.2005.1727.)

Hebert, P. D. N., Cywinska, A., Ball, S. L. & deWaard, J. R. 2003 Biological identifications through DNA barcodes. *Proc. R. Soc. B* **270**, 313–321. (doi:10.1098/rspb.2002.2218.)

Hebert, P. D. N., Penton, E. H., Burns, J. M., Janzen, D. H. & Hallwachs, W. 2004 Ten species in one: DNA barcoding reveals cryptic species in the neotropical skipper butterfly *Astraptes fulgerator*. *PNAS* **101**, 14 812–14 817. (doi:10.1073/pnas.0406166101.)

Janzen, D. H. 1993 Taxonomy: universal and essential infrastructure for development and management of tropical wildland biodiversity. In *Proceedings of the Norway/UNEP Expert Conference on Biodiversity, Trondheim, Norway* (ed. O. T. Sandlund & P. J. Schei), pp. 100–113. Trondheim, Norway: NINA.

Janzen, D. H. 2000 Costa Rica's Area de Conservación Guanacaste: a long march to survival through non-damaging biodevelopment. *Biodiversity* **1**, 7–20.

Janzen, D. H. 2003 How polyphagous are Costa Rican dry forest saturniid caterpillars? In *Arthropods of tropical forests. Spatio-temporal dynamics and resource use in the canopy* (ed. Y. Basset, V. Novotny, S. E. Miller & R.L Kitching), pp. 369–379. Cambridge, UK: Cambridge University Press.

Janzen, D. H. 2004*a* Setting up tropical biodiversity for conservation through non-damaging use: participation by parataxonomists. *J. Appl. Ecol.* **41**, 181–187. (doi:10.1111/j.1365-2664.2004.00879.x.)

Janzen, D. H. 2004*b* Now is the time. *Phil. Trans. R. Soc. B* **359**, 731–732. (doi:10.1098/rstb.2003.1444.)

Janzen, D. H. & Hallwachs, W. 2005 Philosophy, navigation and use of a dynamic database ('ACG Caterpillars SRNP') for an inventory of the macrocaterpillar fauna, and its food plants and parasitoids, of the Area de Conservacion Guanacaste (ACG), northwestern Costa Rica (http://janzen.sas.upenn.edu).

Lemaire, C. 2002 *The Saturniidae of America. Hemileucinae*. Germany: Goecke & Evers, Keltern.

Prendini, L., Hanner, R. & DeSalle, R. 2002 Obtaining, storing and archiving specimens for molecular genetic research. In *Techniques in molecular systematics and evolution. Methods and tools in biosciences and medicine* (ed. R. DeSalle, G. Giribet & W. Wheeler), pp. 176–248. Basel, Switzerland: Birkhauser.

Simmons, R. E., Du Plessis, M. A. & Hedderson, T. A. J. 2005 Seeing the woodhoopoe for the trees: should we abandon Namibia's violet woodhoopoe *Phoeniculus damarensis* as a species? *Ibis* **147**, 222–224. (doi:10.1111/j.1474-919x.2005.00385.x.)

Vaglia, T. & Haxaire, J. 2003 Description d'un nouveau Sphingidae du Costa Rica *Xylophanes letiranti* (Lepidoptera: Sphingidae). *Lambillionea* **103**, 287–290.

Wuster, W., Ferguson, J. E., Quijada-Mascareñas, A. & Pook, C. E. 2005 Tracing an invasion: landbridges, refugia, and the phylogeography of the Neotropical rattlesnake (Serpentes: Viperidae: *Crotalus durissus*). *Mol. Ecol.* **14**, 1095–1108. (doi:10.1111/j.1365-294X.2005.02471.x.)

The supplementary Electronic Appendix is available at http://dx.doi.org/10.1098/rstb.2005.1715 or via http://www.journals.royalsoc.ac.uk.