MOLECULAR ECOLOGY
RESOURCES
Molecular Ecology Resources (2012) **12**, 676–685 doi: 10.1111/j.1755-0998.2012.03143.x

Utility of the DNA barcoding gene fragment for parasitic wasp phylogeny (Hymenoptera: Ichneumonoidea): data release and new measure of taxonomic congruence

DONALD L. J. QUICKE,[1,2] M. ALEX SMITH,[3] DANIEL H. JANZEN,[4] WINNIE HALLWACHS,[4] JOSE FERNANDEZ-TRIANA,[3,5] NINA M. LAURENNE,[6] ALEJANDRO ZALDÍVAR-RIVERÓN,[7] MARK R. SHAW,[8] GAVIN R. BROAD,[2] SERAINA KLOPFSTEIN,[9] SCOTT R. SHAW,[10] JAN HRCEK,[11] PAUL D. N. HEBERT,[3] SCOTT E. MILLER,[12] JOSEPHINE J. RODRIGUEZ,[13] JAMES B. WHITFIELD,[14] MICHAEL J. SHARKEY,[15] BARBARA J. SHARANOWSKI,[16]* REIJO JUSSILA,[17] IAN D. GAULD[DECEASED],[2] DOUGLAS CHESTERS[1,2] and ALFRIED P. VOGLER[1,2]

[1]*Division of Biology, Imperial College London SW7 2AZ, UK,* [2]*Department of Entomology, The Natural History Museum, Cromwell Rd, London SW7 5DB, UK,* [3]*Department of Integrative Biology, Biodiversity Institute of Ontario, University of Guelph, Guelph, Ontario, Canada N1G 2W1,* [4]*Department of Biology, University of Pennsylvania, Philadelphia, PA 19104-6018, USA,* [5]*Canadian National Collection of Insects, 960 Carling Ave., Ottawa, Ontario, Canada K1A 0C6,* [6]*Aalto University School of Science, Department of Media Technology, PO Box 15500, FI-00076 Aalto, Finland and University of Helsinki, Department of Computer Science, Finland,* [7]*Colección Nacional de Insectos, Instituto de Biología, Universidad Nacional Autonoma de México, A.P. 70-153, C.P. 04510, México D. F.,* [8]*Honorary Research Associate, National Museums of Scotland, Chambers Street, Edinburgh, EH1 1JF, UK,* [9]*Natural History Museum, Department of Invertebrates, Bernastrasse 15, CH-3005 Bern, Switzerland, and Zoological Institute, Division of Community Ecology, University of Bern, Baltzerstrasse 6, CH-3012 Bern, Switzerland,* [10]*University of Wyoming Insect Museum, Department of Renewable Resources (3354), University of Wyoming, 1000 East University Avenue, Laramie, WY 82071, USA,* [11]*Biology Center, Czech Academy of Sciences and Faculty of Science, University of South Bohemia, Branisovska 31, 37005 Ceske Budejovice, Czech Republic,* [12]*Smithsonian Institution, PO Box 37012, MRC 105 Washington, DC 20013-7012, USA,* [13]*National Center for Ecological, Analysis and Synthesis (NCEAS), University of California, Santa Barbara, 735 State St., Suite 300 Santa Barbara, CA 93101, USA,* [14]*Department of Entomology, 320 Morrill Hall, University of Illinois, 505 S. Goodwin Ave, Urbana, IL 61801, USA,* [15]*Department of Entomology, University of Kentucky, Lexington, KY 40546-0091, USA,* [16]*North Carolina State University, Department of Entomology, Campus Box 7613, Raleigh, NC 27695, USA,* [17]*Krouvintie 92, 21330 Paattinen, Finland*

Abstract

The enormous cytochrome oxidase subunit I (COI) sequence database being assembled from the various DNA barcoding projects as well as from independent phylogenetic studies constitutes an almost unprecedented amount of data for molecular systematics, in addition to its role in species identification and discovery. As part of a study of the potential of this gene fragment to improve the accuracy of phylogenetic reconstructions, and in particular, exploring the effects of dense taxon sampling, we have assembled a data set for the hyperdiverse, cosmopolitan parasitic wasp superfamily Ichneumonoidea, including the release of 1793 unpublished sequences. Of approximately 84 currently recognized Ichneumonoidea subfamilies, 2500 genera and 41 000 described species, barcoding 5′-COI data were assembled for 4168 putative species-level terminals (many undescribed), representing 671 genera and all but ten of the currently recognized subfamilies. After the removal of identical and near-identical sequences, the 4174 initial sequences were reduced to 3278. We show that when subjected to phylogenetic analysis using both maximum likelihood and parsimony, there is a broad correlation between taxonomic congruence and number of included sequences. We additionally present a new measure of taxonomic congruence based upon the Simpson diversity index, the Simpson dominance index, which gives greater weight to morphologically recognized taxonomic groups (subfamilies) recovered with most representatives in one or a few contiguous groups or subclusters.

Keywords: Braconidae, cytochrome oxidase 1, Ichneumonidae, sampling density, Simpson dominance index, taxonomic retention index

Received 7 November 2011; revision received 18 February 2012; accepted 27 February 2012

Correspondence: Donald L. J. Quicke, Fax: +44(0)2075942339; E-mail: d.quicke@imperial.ac.uk

*Present address for Barbara J. Sharanowski: University of Manitoba, Department of Entomology, 214 Animal Science Bldg, Winnipeg, Manitoba, Canada R3T 2N2.

Introduction

Mitochondrial protein-coding genes have played a large role in many attempts at phylogeny reconstruction, either alone or in concert with other genes. They are easily amplified owing to their high copy number, and they often display some highly conserved primer-binding sites. On the other hand, the high rates at which they evolve mean they often show saturation of substitutions. Two gene fragments have played a particularly important role in studies of Hymenoptera phylogeny in recent years, the nuclear ribosomal 28S and the mitochondrial cytochrome oxidase subunit 1 (COI) (Mardulyn & Whitfield 1999; Zaldívar-Riverón *et al.* 2008; Quicke *et al.* 2009). However, both pose their own problems. The former is length variable and thus difficult to align, and how best to cope with ambiguously aligned regions is still under debate (Gillespie *et al.* 2005; Laurenne *et al.* 2006; Zaldívar-Riverón *et al.* 2006). COI, while being highly constrained in the number of first and second codon position sites free to vary (Graybeal 1994), evolves far more rapidly at its 3rd codon position, and there has been much debate over its utility for recovering phylogenies (Klopfstein *et al.* 2010), especially for deeper (older) divergences.

Recently, a vast additional source of mitochondrial COI sequences has become available from Barcode of Life Datasystem (BOLD; HYPERLINK http://www.boldsystems.org; Ratnasingham & Hebert 2007) (Hebert *et al.* 2003; Smith *et al.* 2007, 2008, 2009; Janzen *et al.* 2009) and similar initiatives. The primary aim of these barcoding programmes has been to develop the means to identify species and reveal morphologically cryptic species on the basis of genetic uniqueness and sequence clustering. However, the observed sequence variation also reflects evolutionary history and is mostly neutral or nearly so. It therefore ought also to provide a wealth of phylogenetic and higher-level taxonomic information. In this study, we concentrate on sequences from the large parasitic Hymenoptera superfamily Ichneumonoidea, and ask whether, with the available data, taxa can be assigned accurately to subfamily level based upon their barcoding sequence and whether accuracy is related to the total number of sequences that are available.

The Ichneumonoidea comprises two huge, cosmopolitan families, the Ichneumonidae with 24 000 and the Braconidae with 17 000 described species in 85 generally recognized subfamilies (Yu *et al.* 2005; Quicke *et al.* 2009). The Braconidae in particular have been subjected to numerous phylogenetic analyses in recent years (Quicke & Achterberg 1990; Wharton *et al.* 1992; Whitfield 1992, 2002; Belshaw *et al.* 1998, 2000; Dowton *et al.* 1998, 2002; Whitfield *et al.* 2002; Shi *et al.* 2005; Sharkey *et al.* 2006; Zaldívar-Riverón *et al.* 2006; Pitz *et al.* 2007; Murphy *et al.* 2008; Sharanowski *et al.* 2011) and have become a test case for phylogeny reconstruction in the face of multiple convergent shifts in biology leading to marked conflict between molecular and morphological data sets (Quicke & Belshaw 1999; Whitfield *et al.* 2002). In contrast, the Ichneumonidae have been subject to relatively few molecular or global morphological investigations (Wahl & Gauld 1998; Quicke *et al.* 2005; Laurenne *et al.* 2006; Klopfstein *et al.* 2010), perhaps owing to them generally being considered more taxonomically difficult to identify at subfamily level with potentially higher levels of homoplasy (Gauld & Mound 1982).

Because COI is rapidly evolving and therefore very prone to saturation, it is necessary firstly to assess the degree and range of signal it contains about the membership of clades (i.e. phylogenetic signal at relevant taxonomic levels). With sequences such as COI, phylogenetic signal and therefore the accuracy with which the sequences are likely to be assigned to correct clades (i.e. higher taxonomic groups) is likely to decrease with increasing taxonomic and temporal distance between taxa as substitutions approach saturation. Nevertheless, Kallersjö *et al.* (1999) showed that the 3rd codon positions of such genes, despite approaching saturation, still contain important phylogenetic signal. Such signal is expected to become useful if taxa are sampled sufficiently densely such that globally homoplasious characters become locally informative and if their rate of evolution is more accurately estimated by the application of an appropriate phylogenetic model.

Assessing phylogenetic accuracy is easy with simulation studies (see Purvis & Quicke 1997) but difficult with real data, as the true phylogeny is typically unknown. Assessing success via congruence with independently derived hypotheses offers the easiest solution. Most classifications are based upon morphology, and even though they are probably imperfect representations of phylogeny, taxonomic hierarchies provide a benchmark for assessing the accuracy of molecular trees.

Recently, the taxonomic retention index (tRI) has been employed as a measure of the congruence between molecular phylogenies and a prior, usually informal, morphological classification at family, subfamily, tribe and genus level (Hunt & Vogler 2008). This index is based upon the retention index (RI; Farris 1989) widely used in phylogenetic studies as a measure of how well the data support the tree. The retention index is defined as follows:

$$\mathrm{RI} = \frac{(G - S)}{(G - M)}$$

where G is the maximum number of evolutionary steps that a character requires when fitted parsimoniously on a tree, M is the minimum number that it could require on

any possible tree, and *S* is the observed number of steps on the given tree. The tRI treats the taxonomic placement of each terminal as a 'pseudocharacter' and is calculated as the ensemble retention index (i.e. using the sum of these values for all the taxonomic groups to calculate the RI) of that on the tree topology. It provides a simple measure that varies linearly with the number of clades that an expected taxon is distributed across, with tRI = 1 if an expected group is recovered as monophyletic and tRI = 0 if all its members are dispersed.

Here, we introduce a new measure based on the Simpson dominance index (SDI; Simpson 1949), the taxonomic Simpson dominance index (tSDI), which reflects of the contiguity of clade(s) that a taxon is split over and also, when a clade is split over more than one cluster on a tree, the distribution of the number of terminals among those clades, thus averting the property of the tRI that a single rogue taxon has the same effect as a substantial split. The SDI was calculated using the sample formula version

$$\mathrm{SDI} = \sum_{i=1}^{s} \frac{n_i(n_i - 1)}{N(N-1)}$$

where N is the total number of terminals (species), s is the number of clusters these are distributed across, and n_i is the number of terminals in the ith group. We chose the Simpson dominance index over various other commonly used indices such as the Shannon-Weiner and Margalef ones because it is less affected by sample size (Giavelli *et al.* 1986) and it is also less affected by a few rare entities.

In terms of interpreting whether sequences are likely to be assigned correctly to higher groups, this is important, because, as widely recognized there are numerous errors present in the taxonomic labels given to sequences on databases such as GenBank (e.g. Aliabadian *et al.* 2009; Lukhtanov *et al.* 2009). In many groups, a few individuals may have been misidentified, or sequences lodged on the databases could have been contaminants or paper-trail mix-ups. These will linearly affect the tRI, whereas if the great majority of sequences fall into the correct expected clade, then the tSDI will be less affected.

Materials and methods

Our initial data set included all readily available COI 5′ barcodes for ichneumonoid wasps, that is, in total 4174 sequences as available from GenBank as of 1st January 2010, many new sequences generated by the ongoing International Barcode of Life project and hosted in the Barcode of Life Data System (BOLD), and sequences from the laboratories of various co-authors and other collaborators. All sequences were realigned manually with reference to amino acid translation. Alignment was unambiguous although a few species displayed single amino acid deletions, single amino acid insertions (*Cubus*, *Triclistus* and *Colopotrochia*) and single base deletions (representatives of the Agathidinae). While such single base deletions are indicative of a pseudogene or NUMT, they are retrieved using multiple primer combinations (M. Smith, unpublished) and present in specimens from widely divergent localities displaying host-specific ecologies. In this analysis, we treated the Agathidinae sequences that included the deletions as the CO1 marker for this species, although further work (beyond the scope of this investigation) is required to determine whether these deletions are pseudogenes, particularly as RNA editing is known to correct gene products, leading to functional transcripts (Russell & Beckenbach 2008; Hrcek *et al.* 2011).

Sequences from approximately 670 genera and 74 subfamilies were represented, including all but four generally recognized subfamilies of Ichneumonidae (lacking Adelognathinae, Microleptinae, Nesomesochorinae and Tatogastrinae) and all but four unambiguously assigned subfamilies of Braconidae (lacking Telengaiinae, Ypsistocerinae, Vaepellinae and Dirrhopinae); we also lack sequences from the endemic Australian Trachypetinae and Chilean Apozyginae which currently are ambiguously placed within the Braconidae (Quicke *et al.* 1999). For many species, or putative species based on sequence divergence and other data, numerous individual sequences were available. From these, we selected the most complete sequence (greatest length, fewest ambiguously called bases) or a random sequence among equally complete sequences.

The data set included several extremely well-sampled genera with many very closely related species differing only by a small number of bases in the target gene region. We thus further pruned the data set to include only a single representative from each cluster of identical or near-identical sequences, considering sequence variation of less than four bases as uninformative. This was achieved through the following algorithm: sequences were compared in a pairwise fashion, with pairs considered identical or near-identical when differing at no more than four bases at unambiguous sites; in other words, sites scored with ambiguity codes [mrwsykvhdbn-] were not included in the string comparisons, so a site containing an 'A' in one of the pairs and 'N' in the other, would be considered identical. Both members of the identical pair were scored for sequence completeness, each additional unambiguously scored base gaining a sequence score of 1. The scoring took the full nucleotide code into account. For example, a sequence with a 'C' at a given position would be considered one base longer than a sequence with a 'Y' (i.e. C or T) at the homologous site. The lower

scoring sequences were then discarded, or in cases where two sequences were identical at shared sites and were also of the same length, the first sequence in the file is retained. After the removal of identical and near-identical sequences, the 4168 Ichneumonoidea sequences were reduced to 3278 (1890 braconid and 1388 ichneumonid sequences, representing 671 genera and 74 subfamilies).

Tree searches

Tree searches were performed using maximum parsimony (MP) and maximum likelihood (ML) algorithms. For MP trees, we used the standard driven TNT search (Goloboff *et al.* 2003), utilizing the constrained and random sectoral search and tree fusing algorithms (TNT command line option: *xmult = replication 5 hits 3 autoconst 1 level 5 multiply css rss fuse 3*). ML trees were built using RAxML (v7.2.5) (Stamatakis *et al.* 2005) using the GTRCAT nucleotide model and four categories for rate variation among sites. The parameter settings chosen for both tree search programmes were based on numerous preliminary runs, and whilst with such a large data set can never guarantee yielding the most likely or parsimonious trees, produce robust searches such that resulting trees are likely to have optimality scores close to the true optima. Indeed, carrying out time-limited searches with TNT showed that even far less thorough searches yielded very similar trees (data not shown). For our purpose, if expected clades are recovered in slightly suboptimal trees, then confidence in the correct taxonomic placement of unknown sequences should be conservative.

Taxonomic congruence scores

Trees were scored for taxonomic congruence of each genus and subfamily using tRI and tSDI. Both indices were calculated using a taxonomic presence/absence matrix. We scored one character for each taxonomic group, with each terminal taxon coded as 1 or 0, depending on whether it belongs to the taxonomic group. The obtained taxonomic matrix was then used as input for PAUP* (Swofford 2001) for the calculation of the parsimony retention index.

Taxonomic Simpson dominance indexes were calculated individually for each taxon using a script written in Perl (available in Table S1, Supporting information). Character states (0 or 1) were then optimized on the tree using the Fitch (1970) parsimony algorithm. With all branches assigned to either state for each taxonomic group, the number of subgroups into which a named taxon had been split and the numbers of terminals in each of those subgroups were inferred (Fig. 1). The

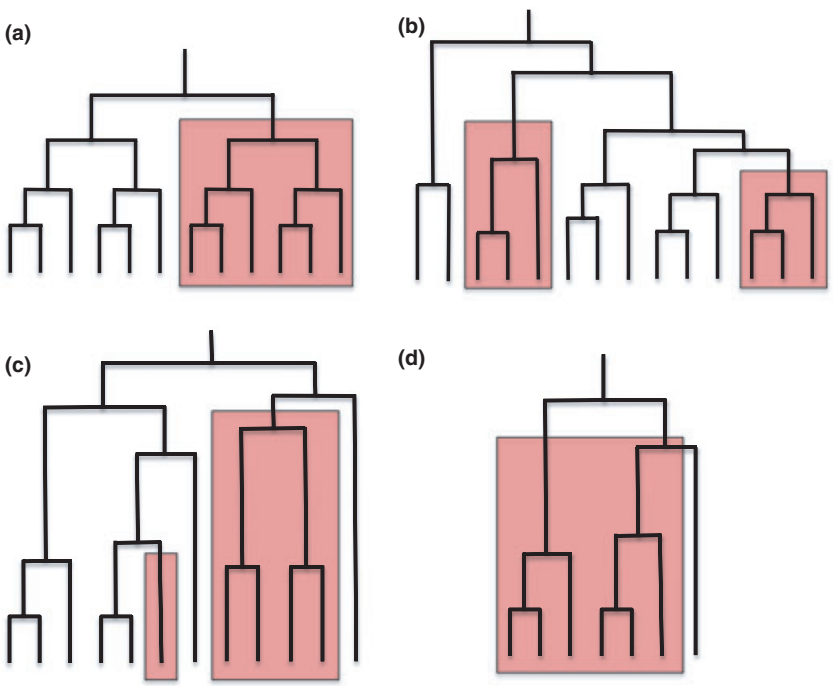

Fig. 1 Illustrating the behaviour of two methods of scoring taxonomic congruence, taxonomic retention index (tRI) and taxonomic Simpson dominance index (tSDI), on four phyletic patterns of a given taxon (here shown as a red box). Branches belonging to the scored taxon are shown as shaded. The monophyletic pattern in (a) gives tSDI of 1.0 and tRI of 1.0. (b) shows the taxon as polyphyletic with two equally sized groups, giving tSDI of 0.4 and tRI of 0.8. In (c), the taxon is also polyphyletic, but of unequally sized groups. Here, the taxon scores tSDI is 0.67, but the tRI remains 0.8. The paraphyletic group in (d) [of five sequences with one invader] gives tSDI of 1.0 and tRI of 0.8.

number of sequences in each subgroup was used for the calculation of tSDI.

For example, if in a 100-taxon tree a taxon with 11 terminals formed two separate lineages, one of one terminal, the other with ten, the tRI would be 0.9, the same as if they were recovered in groups of five and six terminals. In contrast, the tSDI would be 0.8347 in the first case but 0.504 in the second, thus giving a relatively higher value to the case where most of the terminals formed a monophyletic group. The tSDI and tRI respond differently to paraphyly. The former treats a single contiguous set of terminals as a single entity irrespective of it being paraphyletic with respect to other groups, whereas tRI effectively responds linearly to the number of instances of paraphyly irrespective of the coherence of the group on the tree.

Sequences that were unidentified to genus level (353 terminals) were pruned from the resulting trees so as not to indicate lack of monophyly when the unidentified terminal might actually be a member of the relevant taxon during the calculations of tRI and tSDI.

If ambiguity was encountered during the calculation of the monophyly score, we excluded the corresponding data point. For example, different ancestral state reconstruction methods (i.e. Acctran vs. Deltran) may disagree when assigning branches to taxa. In these cases, the count of the number of clades belonging to a specific taxon would be dependent on the method used, and we therefore discarded this score. These cases of ambiguous polyphyly were most common in some of the densely sampled taxa (Ophioninae, Cryptinae and Rogadinae).

Additional programming was carried out using R (R Development Core 2007).

Results

The complete data set was subjected to tree searches, and the ML tree obtained is presented in Fig. 2 in which different subfamilies are colour coded. This shows that most sequences assigned to any given subfamily were usually recovered largely as contiguous units and that separation at family level was complete. Trees were scored for taxonomic congruence using both tRI and tSDI at subfamily level and at genus level. The congruence scores for subfamily level are detailed in Table 1 and the results for subfamily and genus levels summarized in Table 2.

Whilst there was a broad significant correlation between number of sequences representing a subfamily and the tRI for both likelihood and parsimony trees (linear models, likelihood: $F = 9.472$, d.f. = 1, 57, $P < 0.005$; parsimony: $F = 9.807$, d.f. = 1, 57, $P < 0.005$), there was no significant relationship with the tSDI. Twelve subfamilies, mostly represented by relatively few sequences, were always recovered as monophyletic, and three subfamilies, the Labeninae, Pedunculinae and Rhyssinae, were never recovered as monophyletic with either likelihood or parsimony-based analyses (these three being represented by very few sequences in the data set: 2, 2 and 3, respectively). Removing the subfamilies that were always recovered as monophyletic, and therefore might be expected to share many apomorphies, from the analyses gives highly significant positive correlations of both tRI and tSDI (linear models, all P values <0.001) with numbers of included sequences per subfamily for both likelihood and parsimony trees.

The tRI scores were significantly higher for the ML tree than on the parsimony tree for the Braconidae (paired t-test 3.3193, d.f. = 33, $P < 0.005$) but comparisons for tRI of the Ichneumonidae, and tSDI for both families were nonsignificant.

With one exception (Ichneumonidae), the proportion of taxonomic groups recovered as monophyletic was higher for the ML than for the MP trees. Recovery of monophyletic groups was lower on the subfamily compared to genus level. Whereas tRI scores were higher at subfamily compared to genus level in the Braconidae, they were approximately the same at both levels for the Ichneumonidae. In contrast, tSDI scores were more or less similar at both levels in the Braconidae, and they were markedly lower at subfamily level in the Ichneumonidae.

Discussion

The recovery of several morphologically recognized subfamilies as monophyletic when the whole data set is analysed, including several represented by a considerable diversity of included genera, suggests that the barcoding gene region has the potential to place many hard to recognize taxa correctly to subfamily and in many cases also to genus level. With the parasitic Hymenoptera, this may be especially important because it will enable data on host groups and ranges of larger taxonomic entities to be estimated from larval stages (i.e. ones found in association with hosts in the field) when identified adults of the same species are not available or have yet to be sequenced.

Using recovery of taxonomic grouping as an indicator of accuracy, we see approximately one-third of subfamilies and just under half of genera were recovered as monophyletic from analysis of the full data set in the ML trees. One limitation of the current study is the relative incompleteness of taxonomic work on the Ichneumonidae (Quicke *et al.* 2009), with several traditional subfamilies appearing as either polyphyletic or paraphyletic based on combined morphological and 28S rDNA sequence data analyses. Nevertheless, of the remaining well-supported groups, the COI trees still fail to recover

Fig. 2 Circular maximum likelihood phylogram from analysis of the whole data set, differentially colour coded for all major subfamilies and with unassigned taxa coloured grey.

many as monophyletic, and therefore, using parsimony or likelihood, any COI sequence of unknown subfamily level identity might currently be misplaced.

Several subfamilies within the Braconidae and Ichneumonidae have consistently been recovered as nonmonophyletic, but when the taxonomic congruence measures are examined, some of these have high scores. In such cases, the most likely explanation is that some sequences have been incorrectly assigned to them, through either morphological misidentification or laboratory level mix-ups. In a few cases, again especially within the Ichneumonidae, the exact limits of some subfamilies are not known, and some may indeed be poly- or paraphyletic as currently constituted. Recovery of sequences within the group/clade comprising the vast majority of the sequences in these cases is likely therefore to constitute a relatively high degree of confidence about the placement of the unknown taxon.

An important finding of this study is that the broad correlation between taxonomic congruence and number of included sequences for many groups (Fig. 3) which is in agreement with numerous other studies that have more generally shown that phylogenetic accuracy is increased by adding more taxa (Poe 1998; Rannala *et al.* 1998; Zwickl & Hillis 2002). In the case of the barcoding CO1 region, our results suggest that, excluding a few cases where clades have highly characteristic sequences (see Table 1; subfamilies with few representatives but with tRI and tSDI = 1), only by having a large reference set of correctly associated sequences representing a clade

Table 1 Taxonomic congruence scores of subfamilies for maximum likelihood and an exemplar maximum parsimony tree based upon analysis of entire data set (subfamilies represented by only one sequence omitted)

Subfamily	No. of genera represented	No. of sequences	Likelihood		Parsimony	
			tSDI	tRI	tSDI	tRI
Braconidae						
Acampsohelconinae	1	5	0.6	0.75	0.60	0.75
Agathidinae	>1	10	1.0	1.0	1.0	0.67
Alysiinae	18	79	0.95	0.93	0.49	0.91
Amicrocentrinae	1	2	1.0	1.0	1.0	1.0
Aphidiinae	13	56	0.47	0.95	0.85	0.95
Blacinae	2	12	0.32	0.73	0.20	0.64
Brachistinae	9	28	0.38	0.72	0.49	0.64
Braconinae	39	86	1.0	0.99	0.97	0.96
Cardiochilinae	7	13	1.0	0.78	0.62	0.67
Cenocoeliinae	2	2	1.0	1.0	1.0	1.0
Charmontiinae	1	4	1.0	1.0	1.0	1.0
Cheloninae	6	136	0.98	0.99	0.97	0.98
Diospilinae	8	26	0.54	0.72	0.42	0.67
Doryctinae	58	97	0.36	0.94	0.89	0.93
Exothecinae	5	9	0.71	0.83	0.71	0.83
Gnamptodontinae	3	7	1.0	1.0	1.0	1.0
Helconinae	10	31	0.50	0.79	0.22	0.72
Homolobinae	2	9	0.78	0.87	0.78	0.75
Hormiinae	2	9	0.19	0.5	0.19	0.5
Ichneutinae	3	5	1.0	1.0	0.4	0.75
Khoikhoinae	2	5	0.60	0.75	1.0	0.75
Lysiterminae	5	11	0.25	0.57	0.25	0.57
Macrocentrinae	5	37	1.0	1.0	1.0	1.0
Maxfischeriinae	1	7	1.0	1.0	1.0	1.0
Mesostoinae	6	8	1.0	1.0	1.0	1.0
Meteorideinae	1	3	1.0	1.0	1.0	1.0
Meteorinae	14	84	0.97	0.97	0.97	0.93
Microgastrinae	21	408	1.0	0.99	0.51	0.99
Miracinae	1	11	0.82	0.9	0.66	0.8
Opiinae	6	28	0.30	0.80	0.62	0.75
Orgilinae	4	26	1.0	0.95	1.0	0.85
Pambolinae	3	6	0.67	0.6	0.27	0.6
Rhysipolinae	4	6	0.13	0.4	0.13	0.2
Rhyssalinae	5	5	1.0	1.0	0.3	0.5
Rogadinae	35	607	0.91	0.99	0.94	0.99
Ichneumonidae						
Acaenitinae	4	4	0.18	0.33	0.5	0.33
Anomaloninae	12	32	0.68	0.91	0.68	0.91
Banchinae	12	32	1.0	0.97	0.73	0.90
Brachyscleromatinae	4	4	1.0	1.0	1.0	1
Campopleginae	20	187	1.0	0.99	1.0	0.97
Claseinae	1	2	1.0	1.0	1.0	1.0
Cremastinae	5	33	1.0	1.0	1.0	1.0
Cryptinae	>99	182	0.89	0.91	0.61	0.83
Ctenopelmatinae	35	73	0.83	0.89	0.91	0.82
Diplazontinae	11	56	0.69	0.96	1.0	0.93
Hybrizontinae	2	4	1	1	1	1
Ichneumoninae	48	97	0.77	0.86	0.72	0.77
Labeninae	2	2	0	0	0	0
Lycorininae	1	5	1	1	1	1
Mesochorinae	3	108	0.96	0.97	0.94	0.97
Metopiinae	11	33	0.37	0.88	0.27	0.81

Table 1 (Continued)

Subfamily	No. of genera represented	No. of sequences	Likelihood		Parsimony	
			tSDI	tRI	tSDI	tRI
Ophioninae	11	150	0.99	0.99	0.99	0.99
Orthocentrinae	15	112	0.49	0.89	0.34	0.87
Pedunculinae	2	2	0	0	0	0
Pimplinae	21	46	0.45	0.60	0.35	0.51
Poemeniinae	3	4	0.33	0	0.33	0.5
Rhyssinae	3	3	0	0	0	0
Tersilochinae	7	23	0.40	0.78	0.4	0.78
Tryphoninae	14	39	0.67	0.69	0.72	0.63

tSDI, taxonomic Simpson dominance index; tRI, taxonomic retention index.

Table 2 Summary of monophyly and taxonomic congruence measures of morphologically defined genera and subfamilies for maximum likelihood (ML) and maximum parsimony (MP) analyses of entire data set

	Method	Proportion of taxa recovered as monophyletic		Mean tRI		Mean tSDI	
		Subfamily	Genus	Subfamily	Genus	Subfamily	Genus
Braconidae	ML	0.379	0.430	0.868	0.742	0.792	0.759
	MP	0.276	0.398	0.801	0.702	0.717	0.729
Ichneumonidae	ML	0.278	0.5	0.704	0.713	0.648	0.765
	MP	0.278	0.443	0.719	0.692	0.642	0.747

tSDI, taxonomic Simpson dominance index; tRI, taxonomic retention index.

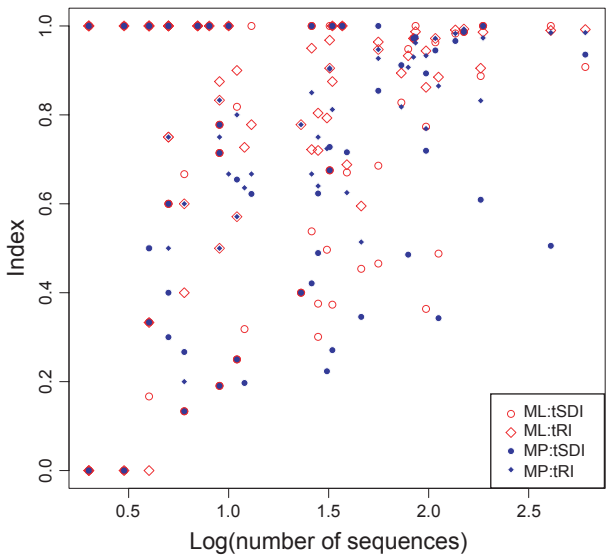

Fig. 3 Relationships between total number of sequences analysed for each subfamily of wasp and two measures of taxonomic congruence (taxonomic retention index and taxonomic Simpson dominance index) on the maximum likelihood tree and an arbitrary most parsimonious tree.

can identify an unknown sequence to that clade be achieved with high certainty. For example, with very well-represented subfamilies such as the Microgastrinae and Rogadinae, it is likely that almost any sequence belonging to a member of either of these will associate in MP or ML analysis with the existing cluster or clusters. In contrast, for relatively poorly represented subfamilies, such as members of the helconine complex within the Braconidae (e.g. Helconinae, Diospilinae, Brachistinae, Blacinae, Orgilinae, Homolobinae), and the Cryptinae, Ichneumoninae, Ctenopelmatinae, Pimplinae, within the Ichneumonidae, the highly interspersed natures of their recovered clades on the best trees (see Fig. 2) suggest that the available CO1 gene fragments are currently too divergent to permit reliable subfamily placements of sequences of unknown origin unless they are very closely related to identified representatives. Therefore, before the available barcoding databases can be used reliably to place the great majority of ichneumonoids to subfamily or tribe level, sampling density will need to be greatly increased for particular, currently under-represented, groups. Finally, this analysis points to the presence in the analysed data set of a few wrongly labelled sequences

which in both ML and MP analyses fall conspicuously among other homogeneous clusters of taxa.

Acknowledgements

We would like to thank the following people who kindly provided additional specimens for sequencing: Andy Deans (North Carolina State University, Raleigh), Brian Fisher and Robert Zuparko (California Academy of Sciences, San Francisco), Kees van Achterberg (Leiden). This research was supported by specimens and barcodes obtained under the following grants: NSF grants BSR 9024770 and DEB 9306296, 9400829, 9705072, 0072730, and 0515699 to D.H. Janzen (as described in Janzen *et al*. 2009); NSF grant (DEB-0841885) to G. Weiblen; NSF grant DEB-0542864 (TIGER Project) to M. J. Sharkey; Grant Agency of the Czech Republic grant 206/09/0115 to V. Novotny; NERC grant NDC519583 to D.L.J. Quicke (and A. Purvis); CONABIO (HB033), CONACyT grants (Red Temática de Códigos de Barra de la Vida, Proyecto Ciencia Básica CONACyT no. 511) to AZR; MAS was supported by Canadian Barcode of Life Research Network from Genome Canada through the Ontario Genomics Institute and an NSERC Discovery Grant. D. Chesters was funded by NERC Case studentship NER/S/A/2006/14013.

References

Aliabadian M, Kaboli M, Nijman V, Vences M (2009) Molecular identification of birds: performance of distance-based DNA barcoding in three genes to delimit parapatric species. *PLoS One*, **4**, e4119.

Belshaw R, Fitton MG, Herniou E, Gimeno C, Quicke DLJ (1998) A phylogenetic reconstruction of the Ichneumonoidea (Hymenoptera) based on the D2 variable region of 28S ribosomal DNA. *Systematic Entomology*, **23**, 109–123.

Belshaw R, Dowton M, Quicke DLJ, Austin AD (2000) Estimating ancestral geographic distributions: a Gondwanan origin for the aphid parasitoids. *Proceedings of the Royal Society, London B*, **267**, 491–496.

Dowton M, Austin AD, Antolin MF (1998) Evolutionary relationships among the Braconidae (Hymenoptera: Ichneumonoidea) inferred from partial 16S rDNA gene sequences. *Insect Molecular Biology*, **7**, 129–150.

Dowton M, Belshaw R, Austin AD, Quicke DLJ (2002) Simultaneous molecular and morphological analysis of braconid relationships (Insecta: Hymenoptera: Braconidae) indicates independent mt-tRNA gene inversions within a single wasp family. *Journal of Molecular Evolution*, **54**, 210–226.

Farris JS (1989) The retention index and the rescaled consistency index. *Cladistics*, **5**, 417–419.

Fitch WM (1970) Distinguishing homologous from analogous proteins. *Systematic Zoology*, **19**, 99–113.

Gauld ID, Mound LA (1982) Homoplasy and the delineation of holophyletic genera in some insect groups. *Systematic Entomology*, **7**, 73–86.

Giavelli G, Rossi O, Sartore F (1986) Comparative evaluation of four species diversity indices related to two specific ecological situations. *Field Studies*, **6**, 429–438.

Gillespie JJ, Yoder MJ, Wharton RA (2005) Predicted secondary structure for 28S and 18S rRNA from Ichneumonoidea (Insecta: Hymenoptera: Apocrita): impact on sequence alignment and phylogeny estimation. *Journal of Molecular Evolution*, **61**, 114–137.

Goloboff P, Farris JS, Nixon K (2003) *TNT: Tree Analysis Using New Technology version 1.1*. [Program and documentation]. The authors, Tucumán, Argentina. Available from http://www.zmuc.dk/public/phylogeny.

Graybeal A (1994) Evaluating the phylogenetic utility of genes: a search for genes informative about deep divergences among vertebrates. *Systematic Biology*, **43**, 174–193.

Hebert PDN, Cywinska A, Ball SL, deWaard JR (2003) Biological identifications through DNA barcodes. *Proceedings of the Royal Society, London B*, **270**, 313–321.

Hrcek J, Miller SE, Quicke DLJ, Smith MA (2011) Molecular detection of trophic links in a complex insect host-parasitoid food web. *Molecular Ecology Resources*, **11**, 786–794.

Hunt T, Vogler AP (2008) A protocol for large-scale rRNA sequence analysis: towards a detailed phylogeny of Coleoptera. *Molecular Phylogenetics and Evolution*, **47**, 289–301.

Janzen DH, Hallwachs W, Blandin P *et al*. (2009) Integration of DNA barcoding into an ongoing inventory of complex tropical biodiversity. *Molecular Ecology Resources*, **9**(Suppl. 1), 1–26.

Kallersjö M, Albert VA, Farris JS (1999) Homoplasy increases phylogenetic structure. *Cladistics*, **15**, 91–93.

Klopfstein S, Kropf C, Quicke DLJ (2010) An evaluation of phylogenetic informativeness profiles and the molecular phylogeny of Diplazontinae (Hymenoptera, Ichneumonidae). *Systematic Biology*, **59**, 226–241.

Laurenne NM, Broad GR, Quicke DLJ (2006) Direct optimization and multiple alignment of 28S D2-3 rDNA sequences: problems with indels on the way to a molecular phylogeny of the cryptine ichneumon wasps (Insecta: Hymenoptera). *Cladistics*, **22**, 442–473.

Lukhtanov VA, Sourakov A, Zakharovs EV, Hebert PDN (2009) DNA barcoding Central Asian butterflies: increasing geographical dimension does not significantly reduce the success of species identification. *Molecular Ecology Resources*, **9**, 1302–1310.

Mardulyn P, Whitfield JB (1999) Phylogenetic signal in the COI, 16S, and 28S genes for inferring relationships among genera of Micrograstrinae (Hymenoptera; Braconidae): evidence of a high diversiWcation rate in this group of parasitoids. *Molecular Phylogenetics and Evolution*, **12**, 282–294.

Murphy N, Banks JC, Whitfield JB, Austin AD (2008) Phylogeny of the microgastroid complex of subfamilies of braconid parasitoid wasps (Hymenoptera) based on sequence data from seven genes, with an improved estimate of the time of origin of the lineage. *Molecular Phylogenetics and Evolution*, **47**, 378–395.

Pitz KM, Dowling AP, Sharanowski BJ, Boring CA, Seltmann KC, Sharkey MJ (2007) Phylogenetic relationships among the Braconidae (Hymenoptera: Ichneumonoidea): a reassessment of Shi *et al*. (2005). *Molecular Phylogenetics and Evolution*, **43**, 338–343.

Poe S (1998) Sensitivity of phylogeny estimation to taxonomic sampling. *Systematic Biology*, **47**, 18–31.

Purvis A, Quicke DLJ (1997) Phylogeny reconstruction: are the big easy? *Trends in Ecology & Evolution*, **12**, 49–50.

Quicke DLJ, Achterberg Cvan (1990) Phylogeny of the subfamilies of the family Braconidae (Hymenoptera, Ichneumonoidea). *Zoologische Verhandelingen*, **258**, 1–95.

Quicke DLJ, Belshaw R (1999) Incongruence between morphological data sets: an example from the evolution of endoparasitism among parasitic wasps (Hymenoptera: Braconidae). *Systematic Biology*, **48**, 436–454.

Quicke DLJ, Basibuyuk HH, Fitton MG, Rasnitsyn AP (1999) Morphological, palaeontological and molecular aspects of ichneumonoid phylogeny (Hymenoptera, Insecta). *Zoologica Scripta*, **28**, 175–202.

Quicke DLJ, Fitton MG, Broad GR, Crocker B, Laurenne NM, Miah IM (2005) The parasitic wasp genera *Skiapus*, *Hellwigia*, *Nonnus*, *Chriodes* and *Klutiana* (Hymenoptera, Ichneumonidae): recognition of the Nesomesochorinae stat. rev. and Nonninae stat. nov. and transfer of *Skiapus* and *Hellwigia* to the Ophioninae. *Journal of Natural History*, **39**, 2559–2578.

Quicke DLJ, Laurenne NM, Fitton MG, Broad GR (2009) A thousand and one wasps: a 28S rDNA and morphological phylogeny of the Ichneumonidae (Insecta: Hymenoptera) with an investigation into alignment parameter space and elision. *Journal of Natural History*, **43**, 1305–1421.

R Development Core Team (2007) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna,

Austria. ISBN 3-900051-07-0, Available from http://www.R-project.org.
Rannala B, Huelsenbeck JP, Yang Z, Nielsen R (1998) Taxon sampling and the accuracy of large phylogenies. *Systematic Biology*, **47**, 702–710.
Ratnasingham S, Hebert PDN (2007) BOLD : the Barcode of Life Data System (http://www.barcodinglife.org). *Molecular Ecology Notes*, **7**, 355–364.
Russell RD, Beckenbach AT (2008) Recoding of translation in turtle mitochondrial genomes: programmed frameshift mutations and evidence of a modified genetic code. *Journal of Molecular Evolution*, **67**, 682–695.
Sharanowski BJ, Dowling APG, Sharkey MJ (2011) Molecular phylogenetics of Braconidae (Hymenoptera: Ichneumonoidea), based on multiple nuclear genes, and implications for classification. *Systematic Entomology*, **36**, 549–572.
Sharkey MJ, Laurenne NM, Quicke DLJ, Sharanowshi BJ, Murray D (2006) Revision of the Agathidinae (Hymenoptera: Braconidae) with comparisons of static and dynamic alignments. *Cladistics*, **22**, 546–567.
Shi M, Chen XX, Achterberg Cvan (2005) Phylogenetic relationships among the Braconidae (Hymenoptera: Ichneumonoidea) inferred from partial 16S rDNA, 28S rDNA D2, 18S rDNA gene sequences and morphological characters. *Molecular Phylogenetics and Evolution*, **37**, 104–116.
Simpson EH (1949) Measurement of biodiversity. *Nature*, **163**, 688.
Smith MA, Wood DM, Janzen DH, Hallwachs W, Hebert PDN (2007) DNA barcodes affirm that 16 species of apparently generalist tropical parasitoid flies (Diptera, Tachinidae) are not all generalists. *Proceedings of the National Academy of Sciences*, **104**, 4967–4972.
Smith MA, Rodríguez JJ, Whitfield JB *et al.* (2008) Extreme diversity of tropical parasitoid wasps exposed by iterative integration of natural history, DNA barcoding, morphology. and collections. *Proceedings of the National Academy of Sciences*, **105**, 12359–12364.
Smith MA, Fernandez-Triana J, Roughley R, Hebert PDN (2009) DNA barcode accumulation curves for understudied taxa and areas. *Molecular Ecology Resources*, **9**(Suppl.1), 208–216.
Stamatakis A, Ludwig T, Meier H (2005) RAxML-III: a fast program for maximum likelihood-based inference of large phylogenetic trees. *Bioinformatics*, **21**, 456–463.
Swofford DL (2001) *PAUP*: Phylogenetic Analysis Using Parsimony (*and Other Methods)*. Sinauer Associates, Sunderland, Massachusetts.
Wahl DB, Gauld ID (1998) The cladistics and higher classification of the Pimpliformes (Hymenoptera: Ichneumonidae). *Systematic Entomology*, **23**, 265–298.
Wharton RA, Shaw SR, Sharkey MJ *et al.* (1992) Phylogeny of the subfamilies of the family Braconidae (Hymenoptera: Ichneumonoidea): a reassessment. *Cladistics*, **8**, 199–235.
Whitfield JB (1992) The polyphyletic origin of endoparasitism in the cyclostome lineages of Braconidae (Hymenoptera). *Systematic Entomology*, **17**, 273–286.
Whitfield JB (2002) Estimating the age of the polydnavirus/braconid wasp symbiosis. *Proceedings of the National Academy of Science*, **99**, 7508–7513.
Whitfield JB, Mardulyn P, Austin AD, Dowton M (2002) Phylogeny of the Microgastrinae (Hymenoptera: Braconidae) based on 16S, COI and 28S genes and morphology. *Systematic Entomology*, **27**, 337–359.
Yu D, van Achterberg C, Horstmann K (2005) *World Ichneumonoidea 2004. Taxonomy, Biology, Morphology and Distribution*. Taxapad, Vancouver, Canada.
Zaldívar-Riverón A, Mori M, Quicke DLJ (2006) Systematics of the cyclostome subfamilies of braconid parasitic wasps (Hymenoptera: Ichneumonoidea): a simultaneous molecular and morphological Bayesian approach. *Molecular Phylogenetics and Evolution*, **38**, 130–145.
Zaldívar-Riverón A, Belokobylskij SA, Leon-Regagnon V, Briceno-G R, Quicke DLJ (2008) Molecular phylogeny and historical biogeography of the cosmopolitan parasitic wasp subfamily Doryctinae (Hymenoptera: Braconidae). *Invertebrate Systematics*, **22**, 345–363.
Zwickl DJ, Hillis DM (2002) Increased taxon sampling greatly reduces phylogenetic error. *Systematic Biology*, **51**, 588–598.

DQ, DC and AV wrote the paper; DQ prepared figures and invented the tSDI; DC and DQ performed the data analysis and computer programming; other authors variously provided sequence data and/or data archiving and/or significant numbers of specimen identifications as well as contributing comments to earlier drafts.

Data accessibility

Details of the final analysed data set with GenBank and BOLD accessions numbers and the taxonomy utilized are provided as Table S1 (Supporting information). The 1793 newly released sequences have accessions numbers in the range from FJ361239 to JF957050. S1 also contains the collection information for those new sequences generated here and the identifier.

Supporting Information

Additional supporting information may be found in the online version of this article.

Table S1. The supplementary material table provides the taxon identifications (family, subfamily, genus and species), process ID, sample ID, Genbank accession number, and collection locality data.

Please note: Wiley-Blackwell are not responsible for the content or functionality of any supporting information supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.