DNA barcode data accurately identify higher taxa

Jonathan A Coddington, Ingi Agnarsson, Ren-Chung Cheng, Klemen Čandek, Amy Driskell, Holger Frick, Matjaž Gregorič, Rok Kostanjšek, Christian Kropf, Matthew Kweskin, Tjaša Lokovšek, Miha Pipan, Nina Vidergar, Matjaž Kuntner

The use of unique DNA sequences as a method for taxonomic identification is no longer fundamentally controversial, even though debate continues on the best markers, methods, and technology to use. Although both existing databanks such as GenBank and BOLD, as well as reference taxonomies, are imperfect, in best case scenarios “barcodes” (whether single or multiple, organelle or nuclear, loci) clearly are an increasingly fast and inexpensive method of identification, especially as compared to manual identification of unknowns by increasingly rare expert taxonomists. Because most species on Earth are undescribed, a complete reference database at the species level is impractical in the near term. The question therefore arises whether unidentified species can, using DNA barcodes, be accurately assigned to more inclusive groups such as genera and families—taxonomic ranks of putatively monophyletic groups for which the global inventory is more complete and stable. We used a carefully chosen test library of CO1 sequences from 49 families, 313 genera, and 816 species of spiders to assess the accuracy of genus and family-level identifications. We used BLAST queries of each sequence against the entire library and got the top ten hits. The percent sequence identity was reported from these hits (PIdent, range 75-100%). Accurate identification (PIdent above which errors totaled less than 5%) occurred for genera at PIdent values > 95 and families at PIdent values ≥ 91, suggesting these as heuristic thresholds for generic and familial identifications in spiders. Accuracy of identification increases with numbers of species/genus and genera/family in the library; above five genera per family and fifteen species per genus all identifications were correct. We propose that using percent sequence identity between conventional barcode sequences may be a feasible and reasonably accurate method to identify animals to family/genus. However, the quality of the underlying database impacts accuracy of results; many outliers in our dataset could be attributed to taxonomic and/or sequencing errors in BOLD and GenBank. It seems that an accurate and complete reference library of families and genera of life *could* provide accurate higher level taxonomic identifications cheaply and accessibly, within years rather than decades.

DNA barcode data accurately identify higher taxa

Jonathan A. Coddington *[1], Ingi Agnarsson [1,5], Ren-Chung Cheng [2], Klemen Čandek [2], Amy Driskell [1], Holger Frick [3], Matjaž Gregorič [2], Rok Kostanjšek [4], Christian Kropf [3], Matthew Kweskin [1], Tjaša Lokovšek [2], Miha Pipan [2,6], Nina Vidergar [2], Matjaž Kuntner *[1,2]

[1] National Museum of Natural History, Smithsonian Institution, Washington, DC, USA
[2] EZ Lab, Institute of Biology at Research Centre of the Slovenian Academy of Sciences and Arts, Ljubljana, Slovenia
[3] Department of Invertebrates, Natural History Museum Bern, Switzerland
[4] Department of Biology, University of Ljubljana, Slovenia
[5] Department of Biology, University of Vermont, Burlington, VT, USA
[6] Currently at: Department of Biochemistry, University of Cambridge, Cambridge, CB2 1GA, United Kingdom

***Authors for correspondence**
Jonathan A. Coddington, Department of Entomology, National Museum of Natural History, Smithsonian Institution, Washington, D.C. 20013-7012, USA, coddington@si.edu
and
Matjaž Kuntner, Research Centre, Slovenian Academy of Sciences and Arts, Novi trg 2, P.O. Box 306, SI-1001, Ljubljana, Slovenia, kuntner@gmail.com

Funding
This work was made possible by a Swiss Contribution to the enlarged EU grant to M. Kuntner and C. Kropf and the Laboratories of Analytical Biology, National Museum of Natural History, Smithsonian Institution.

Keywords: taxonomic impediment, genus, family, genome, Global Genome Initiative, Smithsonian.

Abstract

The use of unique DNA sequences as a method for taxonomic identification is no longer fundamentally controversial, even though debate continues on the best markers, methods, and technology to use. Although both existing databanks such as GenBank and BOLD, as well as reference taxonomies, are imperfect, in best case scenarios "barcodes" (whether single or multiple, organelle or nuclear, loci) clearly are an increasingly fast and inexpensive method of identification, especially as compared to manual identification of unknowns by increasingly rare expert taxonomists. Because most species on Earth are undescribed, a complete reference database at the species level is impractical in the near term. The question therefore arises whether unidentified species can, using DNA barcodes, be accurately assigned to more inclusive groups such as genera and families—taxonomic ranks of putatively monophyletic groups for which the global inventory is more complete and stable. We used a carefully chosen test library of CO1 sequences from 49 families, 313 genera, and 816 species of spiders to assess the accuracy of genus and family-level identifications. We used BLAST queries of each sequence against the entire library and got the top ten hits. The percent sequence identity was reported from these hits (PIdent, range 75-100%). Accurate identification (PIdent above which errors totaled less than 5%) occurred for genera at PIdent values > 95 and families at PIdent values ≥ 91, suggesting these as heuristic thresholds for generic and familial identifications in spiders. Accuracy of identification increases with numbers of species/genus and genera/family in the library; above five genera per family and fifteen species per genus all identifications were correct. We propose that using percent sequence identity between conventional barcode sequences may be a feasible and reasonably accurate method to identify animals to family/genus. However, the quality of the underlying database impacts accuracy of results; many outliers in our dataset could be attributed to taxonomic and/or sequencing errors in BOLD and GenBank. It seems that an accurate and complete reference library of families and genera of life *could* provide accurate higher level taxonomic identifications cheaply and accessibly, within years rather than decades.

Introduction

Accurate identification of biological specimens has always limited the application of biological data to important societal problems. Obstacles are well-known and difficult: the vast majority of species are undescribed scientifically (Erwin, 1982; May, 1992; Mora et al., 2011); some unknown but large fraction of higher taxa are not monophyletic (Goloboff et al., 2009; Pyron & Wiens, 2011); many species can only be identified if certain life stages are available, e.g. adults (Coddington & Levi, 1991), classical data sources such as morphology imperfectly track species identity; the discipline of taxonomy continues to dwindle; the classical process of taxonomic identification is mostly manual and cannot scale to provide the amounts of data required for real-time decisions such as environmental monitoring, invasive species, climate change, etc.

DNA sequence data potentially can eliminate most of these obstacles. DNA barcoding uses a fragment of the mitochondrial gene cytochrome *c* oxidase subunit I (CO1) as a unique species diagnosis/identification tool in the animal kingdom (Hebert et al., 2003), with analogous single to several locus protocols applied for vascular plants, ferns, mosses, algae and fungi (Saunders, 2005; Kress & Erickson, 2007; Nitta, 2008; Chase & Fay, 2009; Liu et al., 2010;), protists (Scicluna, Tawari & Clark, 2006), and prokaryotes (Barraclough et al., 2009). Due to relative ease and inexpensive sequencing, DNA barcoding is a popular tool in species identification and taxonomic applications (e.g. Doña et al., 2015; Xu et al., 2015), and the method is no longer fundamentally controversial at the species level (Pentinsaari, Hebert & Mutanen, 2014; Lopardo & Uhl, 2014; Čandek & Kuntner, 2015; Anslan & Tedersoo, 2015; Wang et al., 2015).

While most species remain undescribed, the situation is not so dire for larger monophyletic groups such as clades accorded the Linnaean ranks of genus or family. In assessing the state of knowledge about biodiversity, it is important to distinguish between the first scientific discovery of an exemplar of a lineage, and phylogenetic understanding of that lineage. Phylogenetic understanding—both tree topology and consequent taxonomic changes, are research programs with no clear end in sight. Linnaean rank is partially arbitrary, and one expects that the number of higher taxa will probably increase over time as understanding improves. Discovery, however, can have an objective definition: the year of the earliest formal taxonomic description of a member of the lineage or taxonomic group in which it is currently included. By this definition the earliest possible discovery of an animal lineage is 1758 (Linnaeus, 1758), or in the case of spiders, 1757 (Clerck, 1757).

More illuminating are the latest discoveries of lineages with the rank of family within larger clades, because the data tell us something about progress towards broad scale knowledge of biodiversity. The species representing the most recent discovery of a family of birds, for example, is the Broad-billed Sapayoa, *Sapayoa aenigma* Hunt, 1903 (Sapayoaidae). The species representing the most recently discovered mammal family is Kitti's hog-nosed bat, *Craseonycteris thonglongyai* Hill, 1974 (Craseonycteridae). For flowering plants, it is *Gomortega keule* (Molina) Baill, 1972 (Gomertegaceae). For bees, it is *Stenotritus elegans* Smith, 1853 (Stenotritidae). For spiders, a megadiverse and poorly known group, it is *Trogloraptor marchingtoni* Griswold, Audisio & Ledford, 2012 (Trogloraptoridae), but the second most recent discovery of an unambiguously new spider family was in 1955, Gradungulidae (Forster, 1955). Figure 1 illustrates the tempo of first discovery of families for these five well-known clades. At the family level, these curves are essentially asymptotic, implying that science is close to completing the inventory of clades ranked as families for these

large lineages. On the other hand, for Bacteria and Archaea (Figure 1), as one would expect, the curve is not asymptotic at all but sharply increasing; prokaryote discovery and understanding is obviously just beginning.

In fact, although many new eukaryote families are named every year, the vast majority of these new names result from advances in phylogenetic understanding, not biological discovery of major new forms of life. The last ten years of Zoological Record suggests that roughly 5-10 truly new families are discovered per year.

In the context of the above question—approximate taxonomic identification of organisms using DNA sequences—these data suggest that our knowledge of major clades of life is approaching completion. The Global Genome Initiative (GGI; http:// http://ggi.si.edu/) of the Smithsonian Institution via the GGI Knowledge Portal (http://ggi.eol.org/) has tabulated a complete list of families of life, which total 9,642—on the whole a surprisingly small number. 10,000 barcodes, more or less, seems like a feasible goal. If we were able to assemble a complete database of DNA sequences at the family level, would it suffice to identify any eukaryote on Earth to the family level?

While the literature on species identification success of DNA barcodes comprises thousands of studies, only a few have tested their effectiveness at the level of higher taxonomic units. In the seminal paper on DNA barcodes Hebert et al. (2003) established that animal CO1 sequences can roughly assign taxa to phyla (96% success) or orders (100% success). However, their test was based on a neighbor joining tree-building approach, and it remained unknown if sequence data itself, i.e. percent identity among taxa, can be used in this way. Similarly, Nagy et al. (2012) showed that DNA barcoding in reptiles usually correctly assigned barcodes to species, genus and family. Their approach was phylogenetic: they tested whether including a sequence in tree building rendered the higher group non-monophyletic, which would imply failure. Finally, Wilson et al. (2011) provided a similar tree based test in sphingid moths, and established reliabilities of correct generic and subfamily taxonomic assignments between 74 and 90% using a liberal, and only 66-84% using a strict, tree-based criterion. These authors argued that tree-based methods perform better than sequence comparison methods, but that reliability, of course, depends on the library completeness.

Our project not only contributes original DNA barcode data for Central European spiders, but also works in synergy with the GGI towards a permanent preservation of genomic biodiversity: the formation of a collection of deeply frozen spider tissues and their DNA. We provide: 1) cryo-preserved tissues of reliably identified species of Central European spiders, and their vouchers photographed and deposited in public museums; 2) permanently frozen genomic DNA of these species; 3) publicly accessible DNA barcodes for these species (genetic sequence of cytochrome oxidase I – CO1) as public identification tool (Hebert et al., 2003) to facilitate organism identification, taxonomy, ecology and conservation.

In addition, this project addresses to what extent higher level taxonomic units can be reliably identified using barcodes of unknown spiders, and specifically asks what percent sequence identity in BLAST results is necessary to correctly identify unknown taxa to the Linnaean genus and/or family. Other methods for classification of higher-level taxonomies such as RDP (Wang et al., 2007), UTAX (Edgar, 2010) and MEGAN (Huson et al., 2007) have primarily been developed for studies of microorganisms, using genetic markers for these groups, but less is

known about using the CO1 barcoding gene in metazoans. We examine empirical data from Araneae barcode data to ask what is the percent sequence identity value above which 5% or less of higher level (genus/family) taxonomic identifications are incorrect and the extent to which frequency of correct identifications correlated with the number of taxa in this dataset, as would be expected given the dependence of BLAST on the reference database.

Materials & Methods

Specimen processing and imaging

We used automated and manual sampling methods for collecting spiders in the field in numerous localities in Slovenia and Switzerland. Faunistic and sampling details are published elsewhere (Čandek et al., 2013; see also 2015 corrigendum). Collected spiders were fixed in absolute ethanol immediately after being caught and the ethanol was replaced on the following day. Spiders were frozen at -80°C, same day, or as soon as possible. In the laboratory they were identified, labeled, photographed and processed for DNA extraction and sequencing (Čandek et al., 2013; see also 2015 corrigendum). Voucher specimens (voucher codes starting with 0078) are deposited at National Museum of Natural History, Smithsonian Institution (Washington D.C., USA), with duplicates (voucher codes starting with ARA) at Naturhistorisches Museum der Burgergemeinde Bern (Switzerland) and EZ LAB, ZRC SAZU (Ljubljana, Slovenia).

Voucher images are published along with their barcodes (see Table 1) at http://ezlab.zrc-sazu.si/dna. All original sequences generated by this project have been submitted to BOLD systems, and those that BOLD accepted were also submitted to GenBank (Table 1).

Tissues

After specimen identification and processing, up to four legs (or in the case of very small individuals the whole prosoma) of a spider were removed and stored in fresh absolute ethanol in cryovials. Part of the tissue was used for DNA isolation while the other part remains permanently frozen at -80 °C at GGI facilities. The maintenance and use of these materials abides by the international legal standards and conventions of the biological genetic heritage (The Access and Benefit Sharing agreement as part of the 2010 Nagoya protocol).

Molecular procedures

At Laboratories of Analytical Biology (National Museum of Natural History, Smithsonian Institution, hereafter LAB), specimens were extracted using the AutoGenPrep phenol-chloroform automated extractor (AutoGen). Samples were digested overnight in buffer containing proteinase-k before extraction. At EZ Lab, specimens (codes starting with ARA) were extracted using the Mag MAX™ Express magnetic particle processor Type 700 with DNA Multisample kit (Applied Biosystems, Foster City, CA) following the manufacturer's protocols.

At EZ Lab PCR was carried out using mainly primers LCO1490 and HCO2198 (Folmer et al., 1994). Standard reaction volume was 35 µL containing 2.3 mM $MgCl_2$ (Promega), 0.15 mM each dNTP (Biotools), 0.4 µM of each primer, 0.2 µL 10 mg/mL BSA (Promega), 0.2 µL GoTaqFlexi polymerase (Promega) and 2 µL DNA. PCR cycling conditions were as follows: an initial denaturation step of 2 min at 94° C followed by 35 cycles of 40 sec at 94° C, 1 min at 48°-52° C, 1 min at 72° C, with final extension at 72° C for 3 min. Additional primers were used for

PCR for a few problematic specimens: dgLCO1490 and dgHCO2198 (Meyer & Paulay, 2005) and the reverse primer Chelicerate-R2 (Barrett & Hebert, 2005). Cycling parameters for difficult specimens were: 20 cycles of usual cycling protocol (above) followed by 15 cycles of 1.5 min at 94° C, 1.5 min at 52° C and 2 min at 72° Cm version 5.6.6 (Kearse et al., 2012). EZ Lab PCR products were sequenced at Macrogen Inc. (Amsterdam, Netherlands), and the sequences were aligned, checked for sequencing errors and trimmed to match the barcode region in Geneious Pro version 5.6.6 (Kearse et al., 2012).

At LAB, PCR was carried out using the primer pair LCO1490 (Folmer et al., 1994) and Chelicerate-R2 (Barrett & Herbert, 2005). A 10 µL reaction mix contained 2.5 mM $MgCl_2$, 0.3 µM of each primer, 0.5 mM dNTPs, and 5 units of Biolase DNA polymerase (Bioline). PCR cycling conditions were as follows: 35 cycles of 30 sec at 95° C, 30 sec at 48° C, 45 sec at 72° C. PCR products were cleaned with ExoSAP-IT (Affymetrix), sequenced using Big Dyes (Life Technologies) and run on a 3730xl DNA sequencer (Applied Biosystems). Sequences were examined using Sequencher 5.01 (Gene Codes).

Barcode library

While we targeted 649 bp long DNA barcodes we also submitted (Table 1) 18 shorter fragments (>570 bp) that still satisfy the requirements of The Barcode of Life Data System BOLD systems (Ratnasingham & Hebert 2007). We combined the 297 species barcodes from this study with publically available Araneae sequences from BOLD retrieved 4 December 2013, for a total of 816 species sequences, which formed the test library for this study. Sequences from BOLD were initially included if the sequence length was at least 600 bases and identification was to species. We further filtered and curated the data to exclude sequences whose identification was anonymous or by non-arachnologists, diverged dramatically from all other spider sequences, or for other reasons the sequences were not deemed to be reliable. After having discarded the above, we did not assess the accuracy of every remaining sequence, as it is well known that both BOLD and GenBank contain errors of various kinds, and we wanted our test library to reflect real world conditions. A single sequence was chosen per species from BOLD using these criteria and added to the original sequences from this project, resulting in 816 species representing 313 genera and 49 families (Table 1 and Supplemental Table 2). Eighteen sequences were singletons at the family level; the maximum number of species per family was 224. 157 sequences were singletons at the genus level; the maximum number of species per genus was 34.

The standalone BLAST+ suite 2.2.28 (Altschul et al., 1990; Zhang et al., 2000) was used to create a custom BLAST database from these sequences. Each sequence was then queried against the full set using blastn (MegaBLAST task, minimum e value of 1e-10, maximum of top ten hits other than the hit of the query to itself). For each hit the percent of identical nucleotides in the aligned region (PIdent) was calculated by BLAST. An advantage of using BLAST is the local nature of the alignment hits returned. This will account for differences in sequence lengths in the dataset, which may otherwise affect pairwise identity calculations of complete alignments. Custom Python scripts (GitHub https://github.com/mkweskin/spider-blast) were used to parse the results, removing the match of the query to itself and to score whether hits matched the genus and family of the query sequence or not. Obviously, if the generic identification matched, the family identification also matched; families therefore always match more often than genera.

On the other hand, singleton generic sequences cannot match correctly at the genus level, and, likewise, singleton family sequences cannot match correctly at the family level. We included singletons as targets in order to model more realistically BLAST searches against the BOLD database (many sequences in BOLD are higher level singletons), and also to test more strongly the ability of sequences with two or more species per either genus or family to match correctly. Including 18 singleton family sequences and 157 singleton genus sequences, therefore, increases the probability of misidentification at either ranks and more strongly tests the usefulness of barcodes as supraspecific identification tools.

However, because the 18 unique family sequences must fail at both the family and genus levels, and the 157 unique genus level sequences must fail at the genus level, these necessary failures were not included in the overall assessments of the ability of barcode sequences to provide accurate identifications at supraspecific levels.

Results

The 816 query sequences returned 8159 total hits with one query only returning nine hits and all others ten (Supplemental Table 1). PIdent scores ranged from 75% to 100%. We also examined the length of the sequence matched compared to the entire sequence length. 8114 hits (>99%) matched to 90% or more of the query sequence length indicating that these results represent matches to large portions of the query validating the use of Percent Sequence Identity in the BLAST hits rather than computing the value for a global alignment between sequences. Figure 2 shows the frequency distributions of PIdent values of correct and incorrect identifications at the genus and family rank.

1. 95% of incorrect genus identifications were below PIdent = 95 when all hits for all queries are included, which suggests the latter value as a heuristic threshold to delimit incorrect from correct identifications (for these data). For only the highest rank hits whose PIdent ≥ 95, 98% of genus identifications were correct.
2. 95% of incorrect family identifications were below PIdent = 91 when all hits for all queries are included, which suggests the latter value as a heuristic threshold to delimit incorrect from correct identifications (for these data). For only the highest rank hits whose PIdent ≥ 91, 97% of family identifications were correct.
3. Library accuracy is crucial, but sequencing, labelling, and identification errors are difficult to detect *a priori*. The highest ranked incorrect family identification was *Steatoda grossa* (Theridiidae) to *Meta menardi* (Tetragnathidae), at PIdent = 96. Further study of the *M. menardi* sequence shows that the BOLD record is probably a mislabeled *Steatoda*. The first true incorrect family identification occurs at a PIdent value of 88; the best hit for *Octonoba* (Uloboridae) is *Amaurobius* (Amaurobiidae).
4. For the 136 genera with at least two species in the library, 76% (n=103) best matched congeners. Thirty-three failed, perhaps because sequences were incorrectly identified taxonomically, or the sequence itself may be erroneous, or perhaps due to non-monophyly of genera.
5. The distributions of PIdents for correct family and genus identifications differ significantly from the distributions of incorrect identifications (Figure 2).
6. Plotted against increasing numbers of species/genus, and genera/family, the proportion of top ten PIdent values that exceed the above suggested threshold values increases.

Roughly speaking, 15 species per genus, and 5 genera per family, are sufficient to ensure that best hits represent correct identifications (Figure 3).

Discussion

We show that standard DNA barcodes can accurately identify unknown specimens to genus and family level given sufficient sequence identity and sufficient taxonomic representation in the database. Accurate identification (PIdent above which less than 5% of identifications were incorrect) occurred for genera at PIdent values > 95 and families at PIdent values ≥ 91, suggesting these as heuristic thresholds for generic and familial identifications in spiders (shaded in Figure 2). Accuracy of identification increases with numbers of species/genus and genera/family; above five genera per family and 15 species per genus all identifications were correct (Figure 3).

The accurate identification of specimens remains a critical challenge for megadiverse groups such as arthropods, most other invertebrates, plants, fungi, protists etc. Morphological identification to species, or even more inclusive taxonomic ranks like genera and families, in many cases requires extensive training, and for most groups taxonomic expertise is limited and dwindling—the so called 'taxonomic impediment' (Rodman & Cody, 2003). DNA barcodes have been proposed as convenient tools to overcome this impediment by making identification a purely technical procedure available to any interested researcher or even 'citizen scientists'. However, the accuracy of such a tool strongly depends on the scope and quality of the barcode library (Smit, Reijnen & Stokvis, 2013). Currently available data on databanks like BOLD and GenBank are extensive for some groups, yet the vast majority of species on earth have not yet been barcoded, much less discovered and described taxonomically—each of these tasks is enormous. Even for existing barcoding data, numerous sequences lack taxonomic identification, limiting their utility (e.g. only 58% of Araneae in BOLD are identified to species, and of those many are not correctly identified, as shown in our results; see also Shen, Chen & Murphy, 2013; Blagoev et al., 2016). Therefore, the identification of unknown specimens through blasting against BOLD or GenBank will be inaccurate if the databases lack close hits or contain errors. While the ideal database would allow species-level identification by containing barcodes from expertly identified and vouchered specimens of all species, we hypothesized that rapid surveys of well-known biotas can help quickly to build valuable tools allowing identification of larger clades such as genera and families.

Although we were careful to screen available barcode sequences from BOLD to produce a test library with as few errors as possible, it is certainly possible that errors remained, either due to mistakes in the lab or taxonomic identifications of vouchers. For example, *Meta menardi* (Tetragnathidae) blasted to *Steatoda grossa* (Theridiidae) at PIdent = 96, and BLAST searches on GenBank suggest this *Meta* sequence is actually a *Steatoda*. Likewise, the linyphiids *Agyneta orites* and *Incestophantes frigidus* sequences were identical; one of these records is probably wrong. These sorts of errors bias identifications and limit utility of barcodes. Other examples of identical barcode sequences were all congeners, and therefore are less likely to involve errors but could indicate faults in taxonomy: *Arctosa maculata* and *A. fulvolineata*, *Bolyphantes luteolus* and *B. alticeps*, *Pardosa alacris* and *P. trifrons*, and *Pityohyphantes tacoma* and *P. cristatus*. Likewise, the genus *Neriene* (Linyphiidae) seems non-monophyletic and identifications were thus not accurate.

Conclusions

These results suggest that accurate identification of unknown taxa to the genus and family level is feasible through DNA barcoding. Database quality is crucial. Numbers of potential matches at generic and familial ranks also affect the probability that an unknown sequence will blast best to the correct family or genus. Unlike the inventory of species, biological discovery of family-level clades of life also seems far advanced—few eukaryotic families, apparently, remain to be discovered. Taken together, these results suggest that barcode-targeted sequencing of exemplars from all families of life (and most genera, if possible) should be an important scientific priority. It would enable approximate taxonomic identification of any organism anywhere on Earth by rapid, cheap, purely technical procedures requiring no specialist knowledge—certainly an important milestone in the on-going attempt to discover, classify, and understand the Earth's biota.

Acknowledgements

Portions of the laboratory and/or computer work were conducted in and with the support of the L.A.B. facilities of the National Museum of Natural History.

References

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *Journal of Molecular Biology* 215:403-410.

Anslan S, Tedersoo L. 2015. Performance of cytochrome c oxidase subunit I (COI), ribosomal DNA Large Subunit (LSU) and Internal Transcribed Spacer 2 (ITS2) in DNA barcoding of Collembola. *European Journal of Soil Biology* 69:1-7. DOI: 10.1016/j.ejsobi.2015.04.001.

Barraclough TG, Hughes M, Ashford-Hodges N, Fujisawa T. 2009. Inferring evolutionarily significant units of bacterial diversity from broad environmental surveys of single-locus data. *Biology Letters* 5:425-428. DOI: 10.1098/rsbl.2009.0091.

Barrett RDH, Hebert PDN. 2005. Identifying spiders through DNA barcodes. *Canadian Journal of Zoology-Revue Canadienne De Zoologie* 83:481-491. DOI: 10.1139/z05-024.

Blagoev GA, Dewaard JR, Ratnasingham S, Dewaard SL, Lu L, Robertson J, Telfer AC, Hebert PDN. 2016. Untangling taxonomy: A DNA barcode reference library for Canadian spiders. *Molecular Ecology Resources* 16:325–341. DOI: 10.1111/1755-0998.12444.

Čandek K, Gregorič M, Kostanjšek R, Frick H, Kropf C, Kuntner M. 2013. Targeting a portion of central European spider diversity for permanent preservation. *Biodiversity Data Journal* 1:e980. DOI: 10.3897/BDJ.1.e980.

Čandek K, Gregorič M, Kostanjšek R, Frick H, Kropf C, Kuntner M. 2015. Corrigendum: Targeting a portion of central European spider diversity for permanent preservation. *Biodiversity Data Journal* 3:e4301. DOI: 10.3897/BDJ.3.e4301.

Čandek K, Kuntner M. 2015. DNA barcoding gap: Reliable species identification over morphological and geographical scales. *Molecular Ecology Resources* 15:268–277. DOI: 10.1111/1755-0998.12304.

Chase MW, Fay MF. 2009. Barcoding of plants and fungi. *Science* 325:682-683. DOI: 10.1126/science.1176906.

Clerck C. 1757. *Aranei Suecici, descriptionibus et figuris oeneis illustrati,ad genera subalterna redacti speciebus ultra LX determinati. Svenska Spindlar,uti sina hufvud-slagter undelte samt.*. Stockholmiae: [Publ. not given].

Coddington JA, Levi HW. 1991. Systematics and evolution of spiders (Araneae). *Annual Review of Ecology and Systematics* 22:565-592. DOI: 10.1146/annurev.es.22.110191.003025.

Doña J, Diaz-Real J, Mironov S, Bazaga P, Serrano D, Jovani R. 2015. DNA barcoding and minibarcoding as a powerful tool for feather mite studies. *Molecular Ecology Resources* 15:1216–1225. DOI: 10.1111/1755-0998.12384.

Edgar RC. 2010. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* 26:2460-2461. DOI: 10.1093/bioinformatics/btq461.

Erwin TL. 1982. Tropical forests: Their richness in Coleoptera and other arthropod species. *The Coleopterists Bull* 36(1):74-75.

Folmer O, Black M, Hoeh W, Lutz R, Vrijenhoek R. 1994. DNA primers for amplification of mitochondrial cytochrome c oxidase subunit I from diverse metazoan invertebrates. *Molecular Marine Biology and Biotechnology* 3:294-299.

Forster RR. 1955. A new family of spiders of the sub-order Hypochilomorphae. *Pacific Science* 9:277-285.

Goloboff PA, Catalano SA, Mirande JM, Szumik CA, Arias JS, Kallersjo M, Farris JS. 2009. Phylogenetic analysis of 73 060 taxa corroborates major eukaryotic groups. *Cladistics* 25:211-230. DOI: 10.1111/j.1096-0031.2009.00255.x.

Hebert PDN, Cywinska A, Ball SL, DeWaard JR. 2003. Biological identifications through DNA barcodes. *Proceedings of the Royal Society of London Series B-Biological Sciences* 270:313-321. DOI: 10.1098/rspb.2002.2218.

Huson DH, Auch AF, Qi J, Schuster SC. 2007. MEGAN analysis of metagenomic data. *Genome Research* 17:377-386. DOI: 10.1101/gr.5969107.

Kearse M, Moir R, Wilson A, Stones-Havas S, Cheung M, Sturrock S, Buxton S, Cooper A, Markowitz S, Duran C, Thierer T, Ashton B, Mentjies P, Drummond A. 2012. Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics*, 28, 1647-1649. DOI: 10.1093/bioinformatics/bts199.

Kress WJ, Erickson DL. 2007. A two-locus global DNA barcode for land plants: the coding rbcL gene complements the non-coding trnH-psbA spacer region. *PLoS ONE* 2(6):e508. DOI: 10.1371/journal.pone.0000508.

Linnaeus C. 1758. *Systema Naturae per regna tria naturae, secundum classes, ordines, genera, species cum characteribus differentiis, synonymis, locis. Editio decima, reformata*. Holmiae, tomus 1:1-821: [Publ. not given].

Liu Y, Yan HF, Cao T, Ge XJ. 2010. Evaluation of 10 plant barcodes in Bryophyta (Mosses). *Journal of Systematics and Evolution* 48:36-46. DOI: 10.1111/j.1759-6831.2009.00063.x.

Lopardo L, Uhl G. 2014. Testing mitochondrial marker efficacy for DNA barcoding in spiders: a test case using the dwarf spider genus *Oedothorax* (Araneae: Linyphiidae: Erigoninae). *Invertebrate Systematics* 28:501-521. DOI: 10.1071/IS14017.

413 May RM. 1992. How many species inhabit the earth. *Scientific American* 1992:42-48.
414 Meyer CP, Paulay G. 2005. DNA barcoding: Error rates based on comprehensive sampling.
415 *PLoS Biology* 3(12):e422. DOI: 10.1371/journal.pbio.0030422.
416 Mora C, Tittensor DP, Adl S, Simpson AG, and Worm B. 2011. How many species are there on
417 Earth and in the ocean? *PLoS Biology* 9(8):e1001127. DOI:
418 10.1371/journal.pbio.1001127.
419 Nagy ZT, Sonet G, Glaw F, and Vences M. 2012. First large-scale DNA barcoding assessment of
420 reptiles in the biodiversity hotspot of Madagascar, based on newly designed COI primers.
421 *PLoS ONE* 7(3):e34506. DOI: 10.1371/journal.pone.0034506.
422 Nitta JH. 2008. Exploring the utility of three plastid loci for biocoding the filmy ferns
423 (Hymenophyllaceae) of Moorea. *Taxon* 57:725-736.
424 Pentinsaari M, Hebert PDN, Mutanen M. 2014. Barcoding beetles: A regional survey of 1872
425 species reveals high identification success and unusually deep interspecific divergences.
426 *PLoS ONE* 9(9):e108651. DOI: 10.1371/journal.pone.0108651.
427 Pyron RA, Wiens JJ. 2011. A large-scale phylogeny of Amphibia including over 2800 species,
428 and a revised classification of extant frogs, salamanders, and caecilians. *Molecular*
429 *Phylogenetics and Evolution* 61:543-583. DOI: 10.1016/j.ympev.2011.06.012.
430 Ratnasingham S, Hebert PDN. 2007. BOLD: The Barcode of Life Data System
431 (www.barcodinglife.org). *Molecular Ecology Notes* 7:355-364. DOI: 10.1111/j.1471-
432 8286.2007.01678.x.
433 Rodman JE, Cody JH. 2003. The taxonomic impediment overcome: NSF's partnerships for
434 enhancing expertise in taxonomy (PEET) as a model. *Systematic Biology* 52:428-435.
435 DOI: 10.1080/10635150390197055.
436 Saunders GW. 2005. Applying DNA barcoding to red macroalgae: A preliminary appraisal holds
437 promise for future applications. *Philosophical Transactions of the Royal Society B-*
438 *Biological Sciences* 360:1879-1888. DOI: 10.1098/rstb.2005.1719.
439 Scicluna SM, Tawari B, Clark CG. 2006. DNA barcoding of Blastocystis. *Protist* 157:77-85.
440 DOI: 10.1016/j.protis.2005.12.001.
441 Shen YY, Chen X, Murphy RW. 2013. Assessing DNA barcoding as a tool for species
442 identification and data quality control. *PLoS ONE* 8(2):e57125. DOI:
443 10.1371/journal.pone.0057125.
444 Smit J, Reijnen B, Stokvis F. 2013. Half of the European fruit fly species barcoded (Diptera,
445 Tephritidae); a feasibility test for molecular identification. *ZooKeys* 365:279-305. DOI:
446 10.3897/zookeys.365.5819.
447 Wang Q, Garrity GM, Tiedje JM, Cole JR. 2007. Naïve bayesian classifier for rapid assignment
448 of rRNA sequences into the new bacterial taxonomy. *Applied and Environmental*
449 *Microbiology* 73:5261-5267. DOI: 10.1128/AEM.00062-07.
450 Wang XB, Deng J, Zhang JT, Zhou QS, Zhang YZ, Wu SA. 2015. DNA barcoding of common
451 soft scales (Hemiptera: Coccoidea: Coccidae) in China. *Bulletin of Entomological*
452 *Research* 105(5):545-554. DOI: 10.1017/S0007485315000413.
453 Wilson JJ, Rougerie R, Schonfeld J, Janzen DH, Hallwachs W, Hajibabaei M, Kitching IJ,
454 Haxaire J, Hebert PDN. 2011. When species matches are unavailable are DNA barcodes
455 correctly assigned to higher taxa? An assessment using sphingid moths. *Bmc Ecology*
456 11:18. DOI: 10.1186/1472-6785-11-18.
457 Xu X, Liu F, Chen J, Li D, Kuntner M. 2015. Integrative taxonomy of the primitively segmented
458 spider genus *Ganthela* (Araneae: Mesothelae: Liphistiidae): DNA barcoding gap agrees

with morphology. *Zoological Journal of the Linnean Society* 175:288–306. DOI: 10.1111/zoj.12280.

Zhang Z, Schwartz S, Wagner L, Miller W. 2000. A greedy algorithm for aligning DNA sequences. *Journal of Computational Biology* 7:203-214. DOI: 10.1089/10665270050081478.

FIGURES

Figure 1. Accumulation curve of dates of first discovery (year of first description of a contained species) of families for six major clades of life, 1758-2010.

Figure 2. Frequency distributions of correct and incorrect identifications by percent sequence identity (PIdent) for the top ten hits (all ranks) and for best hits only (top rank) at the genus and family level. Shaded areas include hits where no more than 5% of identifications were incorrect.

Figure 3. Relation between proportion of best sequence identity and numbers of species per genus, and genera per family (heuristic thresholds to delimit incorrect from correct identifications were 95 and 91 for genus and family, respectively).

TABLES

Table 1. Original sequences this project submitted to BOLD and GenBank (only those on GenBank are also publically available on BOLD, for all others, see http://ezlab.zrc-sazu.si/dna/). Legend: MNH, SI = National Museum of Natural History, Smithsonian Institution; EZ LAB = Evolutionary Zoology Lab, ZRC SAZU; NMBE = Naturhistorisches Museum der Burgergemeinde Bern; SVN = Slovenia; CHE = Switzerland; MYS = Malaysia.

See separate Excel file.

SUPPLEMENTS (AVAILABLE ONLINE)

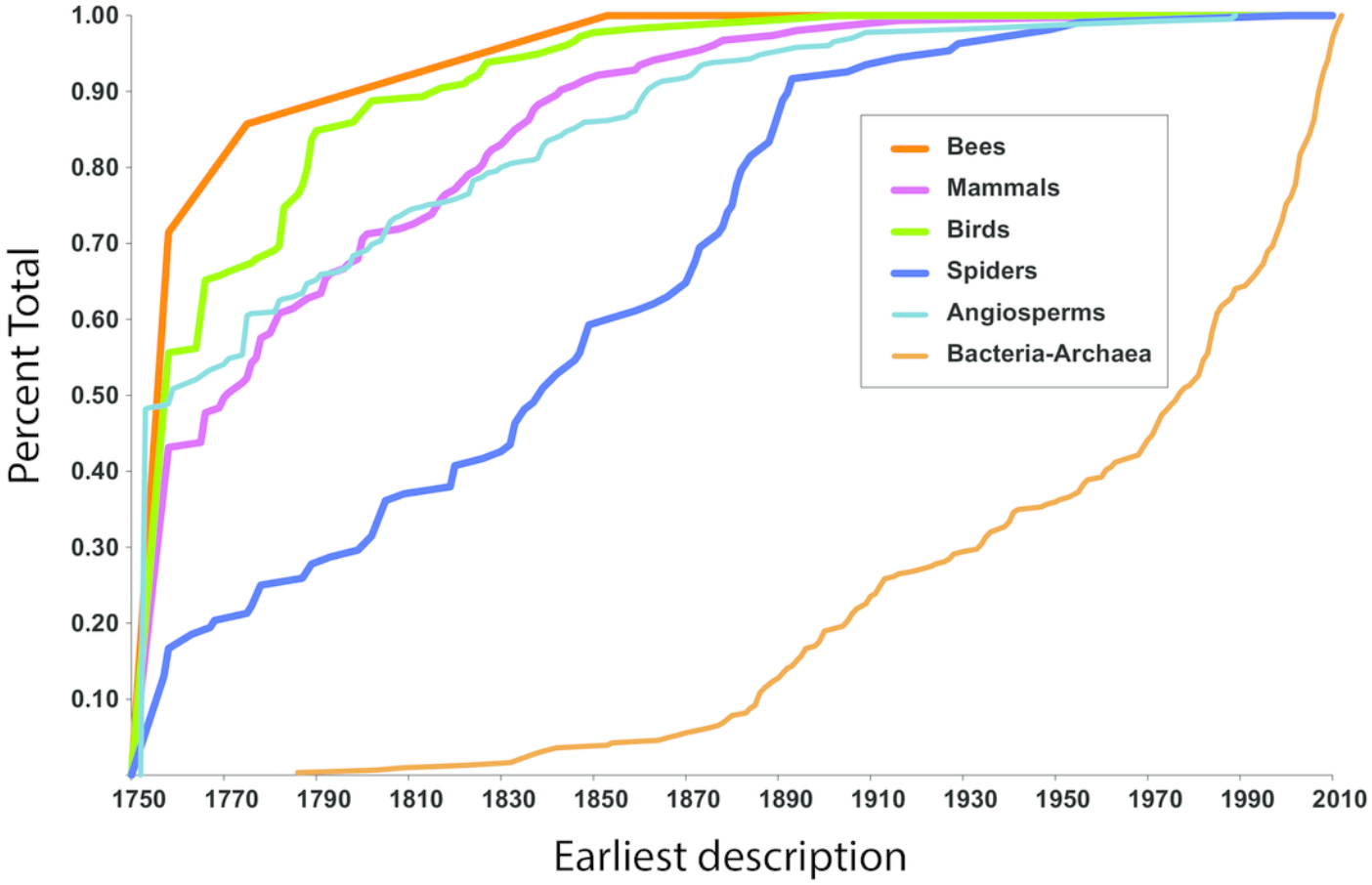
Supplemental Table 1. The results of the barcode matching test.

Supplemental Table 2. The downloaded sequences used in the species comparison.

1

First discovery of major clades of life.

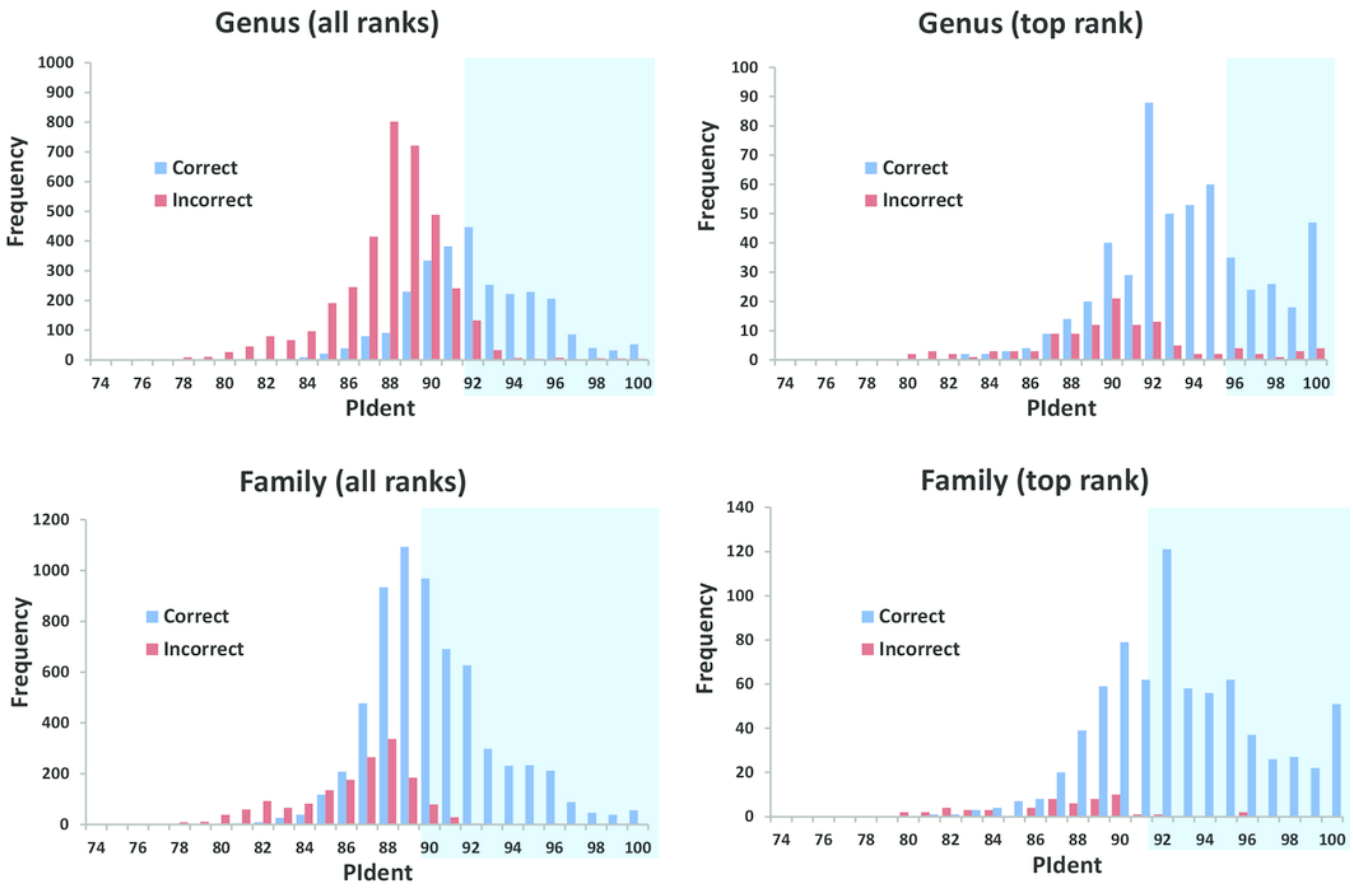
Figure 1. Accumulation curve of dates of first discovery (year of first description of a contained species) of families for six major clades of life, 1758-2010.



2

Results form the barcode matching test.

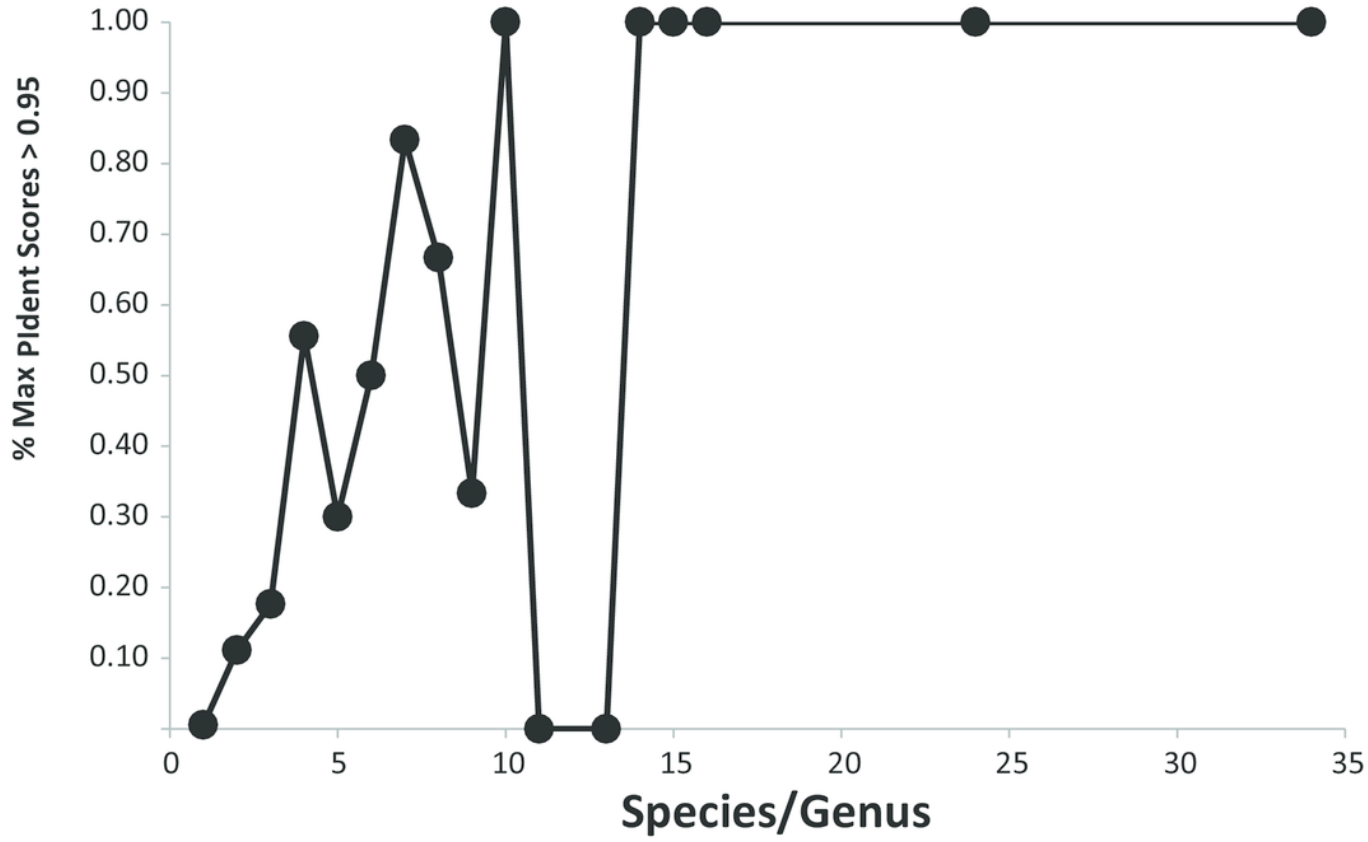
Figure 2. Frequency distributions of correct and incorrect identifications by percent sequence identity (PIdent) for the top ten and/or best hits at the genus and family level. Shaded areas include hits where no more than 5% of identifications were incorrect.

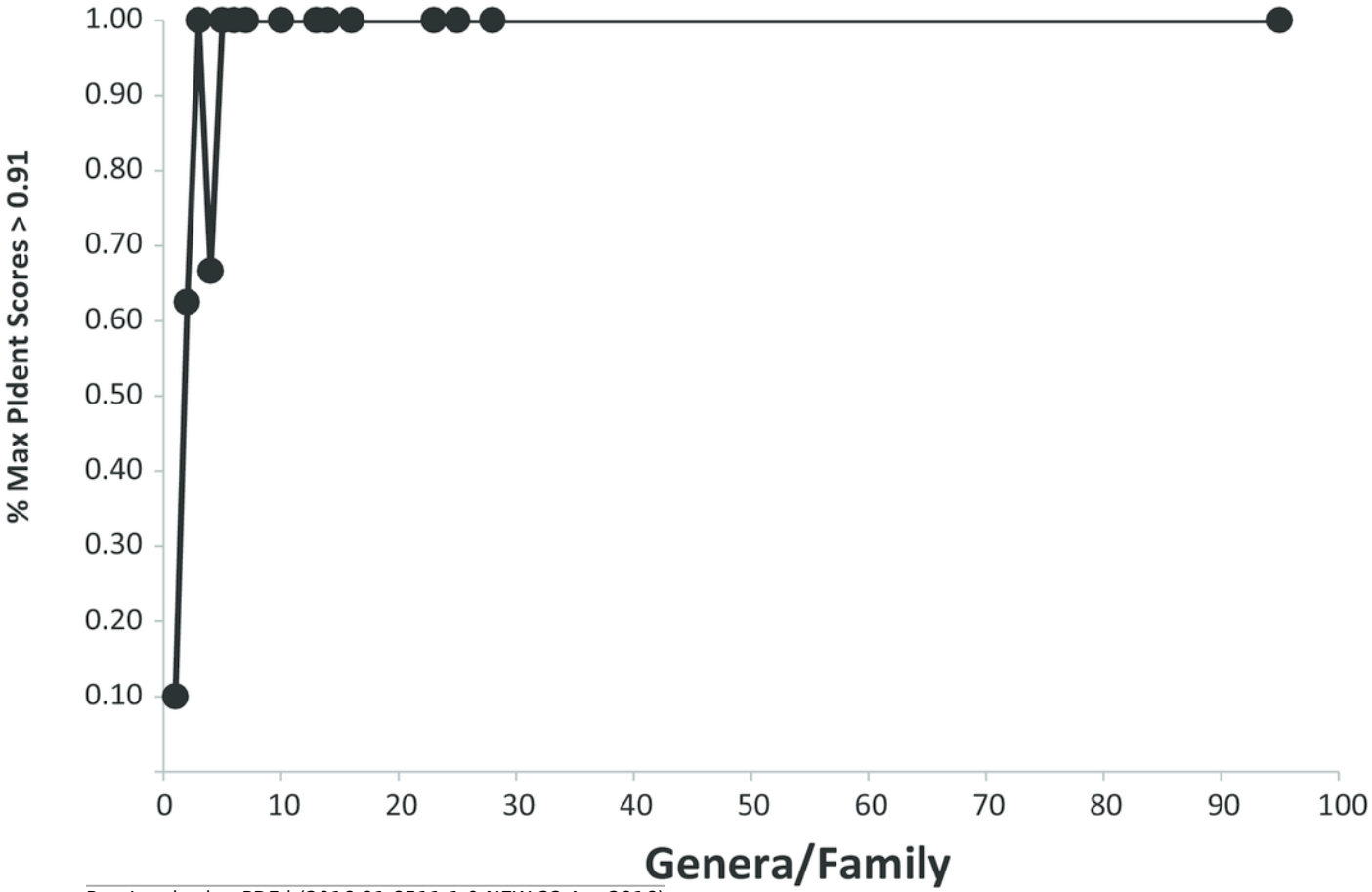


3

Importance of library representation.

Figure 3. Relation between proportion of best sequence identity and numbers of species per genus, and genera per family (thresholds 95 and 91 respectively).


% Max PIdent Scores > 0.95
1.00
0.90
0.80
0.70
0.60
0.50
0.40
0.30
0.20
0.10
0
5
10
15
20
25
30
35
Species/Genus


% Max PIdent Scores > 0.91
1.00
0.90
0.80
0.70
0.60
0.50
0.40
0.30
0.20
0.10
0
10
20
30
40
50
60
70
80
90
100
Genera/Family