DOI: 10.1111/1755-0998.13510

RESOURCE ARTICLE

MOLECULAR ECOLOGY RESOURCES WILEY

A molecular-based identification resource for the arthropods of Finland

Tomas Roslin[1,2] | Panu Somervuo[3] | Mikko Pentinsaari[4] | Paul D. N. Hebert[4] | Jireh Agda[4] | Petri Ahlroth[5] | Perttu Anttonen[6,7] | Jouni Aspi[8] | Gergin Blagoev[4] | Santiago Blanco[4] | Dean Chan[4] | Tom Clayhills[9] | Jeremy deWaard[4] | Stephanie deWaard[4] | Tyler Elliot[4] | Riikka Elo[10,11] | Sami Haapala[12] | Eero Helve[13] | Jari Ilmonen[14] | Petri Hirvonen[15] | Chris Ho[4] | Juhani Itämies[12] | Vladislav Ivanov[8] | Jevgeni Jakovlev[16] | Aino Juslén[17] | Reijo Jussila[18] | Jere Kahanpää[19] | Lauri Kaila[19] | Jari-PekkaKaitila[20] | Ari Kakko[12] | Iiro Kakko[21] | Ali Karhu[22] | Sami Karjalainen[23] | Jostein Kjaerandsen[24] | Janne Koskinen[2,25] | Erkki M. Laasonen[26] | Leena Laasonen[26] | Erkka Laine[27] | Petri Lampila[26] | Valerie Levesque-Beaudin[4] | Liuqiong Lu[4] | Meri Lähteenaro[28,29] | Pekka Majuri[30] | Sampsa Malmberg[13] | Ramya Manjunath[4] | Petri Martikainen[31] | Jaakko Mattila[18] | Jaclyn McKeown[4] | Petri Metsälä[32] | Margarita Miklasevskaja[4] | Meredith Miller[4] | Renee Miskie[4] | Arto Muinonen[33] | Veli-MattiMukkala[34] | Suresh Naik[4] | Nadia Nikolova[4] | Kari Nupponen[13] | Otso Ovaskainen[3,35,36] | Ika Österblad[37] | Lauri Paasivirta[38] | Timo Pajunen[17] | Petri Parkko[39] | Juho Paukkunen[19] | Ritva Penttinen[10,11] | Kate Perez[4] | Jaakko Pohjoismäki[25] | Sean Prosser[4] | Martti Raekunnas[40] | Miduna Rahulan[4] | Meeri Rannisto[17] | Sujeevan Ratnasingham[4] | Pekka Raukko[41] | Aki Rinne[26] | Teemu Rintala[42] | Susana Miranda Romo[4] | Jukka Salmela[43,44] | Juha Salokannel[45] | Riitta Savolainen[3] | Leif Schulman[5,17] | Pasi Sihvonen[17] | Dina Soliman[4] | Jayme Sones[4] | Claudia Steinke[4] | Gunilla Ståhls[17] | Jukka Tabell[46] | Mikko Tiusanen[2] | Gergely Várkonyi[47] | Eero J. Vesterinen[1,48] | Esko Viitanen[13] | Veli Vikberg[49] | Matti Viitasaari[26] | Jussi Vilen[50] | Connor Warne[4] | Catherine Wei[4] | Kaj Winqvist[51] | Evgeny Zakharov[4] | Marko Mutanen[8]

[1]Department of Ecology, Swedish University of Agricultural Sciences, Uppsala, Sweden

[2]Department of Agricultural Sciences, University of Helsinki, Helsinki, Finland

[3]Organismal and Evolutionary Biology Research Programme, University of Helsinki, Helsinki, Finland

[4]Centre for Biodiversity Genomics, University of Guelph, Guelph, ON, Canada

[5]Finnish Environment Institute (SYKE), Helsinki, Finland

[6]Institute of Biology/Geobotany and Botanical Garden, Martin Luther University Halle-Wittenberg, Halle, Germany

Roslin and Somervuo contributed equally to the study.

[7]German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig, Leipzig, Germany
[8]Ecology and Genetics Research Unit, University of Oulu, Oulu, Finland
[9]Parainen, Finland
[10]Zoological Museum, Biodiversity Unit, University of Turku, Turku, Finland
[11]Zoology Unit, Finnish Museum of Natural History, University of Helsinki, Helsinki, Finland
[12]Oulu, Finland
[13]Espoo, Finland
[14]Metsähallitus, Parks & Wildlife Finland, Vantaa, Finland
[15]Porvoo, Finland
[16]Vantaa, Finland
[17]Finnish Museum of Natural History 'Luomus', University of Helsinki, Helsinki, Finland
[18]Paattinen, Finland
[19]Zoology Unit, Finnish Museum of Natural History, University of Helsinki, Helsinki, Finland
[20]Kerava, Finland
[21]Forssa Museum of Natural History, Forssa, Finland
[22]Viinijärvi, Finland
[23]Kirkkonummi, Finland
[24]The Arctic University Museum of Norway, UiT –The Arctic University of Norway, Langnes, Tromsø, Norway
[25]Department of Environmental and Biological Sciences, University of Eastern Finland, Joensuu, Finland
[26]Helsinki, Finland
[27]Jyväskylä, Finland
[28]Division of Systematics, Department of Zoology, Stockholm University, Stockholm, Sweden
[29]Department of Entomology, Swedish Museum of Natural History, Stockholm, Sweden
[30]Liminka, Finland
[31]Koikkala, Finland
[32]Metsäkylä, Finland
[33]Savonlinna, Finland
[34]Kaarina, Finland
[35]Department of Biological and Environmental Science, University of Jyväskylä, Jyväskylä, Finland
[36]Department of Biology, Centre for Biodiversity Dynamics, Norwegian University of Science and Technology, Trondheim, Norway
[37]Korsholm, Finland
[38]Salo, Finland
[39]Kouvola, Finland
[40]Iittala, Finland
[41]Ummeljoki, Finland
[42]Hämeenlinna, Finland
[43]Regional Museum of Lapland, Arktikum, Rovaniemi, Finland
[44]Arctic Centre, University of Lapland, Rovaniemi, Finland
[45]Tampere, Finland
[46]Hartola, Finland
[47]Biodiversity Centre, Finnish Environment Institute SYKE, Kuhmo, Finland
[48]Department of Biology, University of Turku, Turku, Finland
[49]Turenki, Finland
[50]Hämeenkoski, Finland
[51]Turku, Finland

Correspondence

Tomas Roslin, Department of Ecology, Swedish University of Agricultural Sciences, P.O. Box 7044, SE-750 07 Uppsala, Sweden.
Email: tomas.roslin@slu.se

Abstract

To associate specimens identified by molecular characters to other biological knowledge, we need reference sequences annotated by Linnaean taxonomy. In this study, we (1) report the creation of a comprehensive reference library of DNA barcodes for the arthropods of an entire country (Finland), (2) publish this library, and (3) deliver a

new identification tool for insects and spiders, as based on this resource. The reference library contains mtDNA COI barcodes for 11,275 (43%) of 26,437 arthropod species known from Finland, including 10,811 (45%) of 23,956 insect species. To quantify the improvement in identification accuracy enabled by the current reference library, we ran 1000 Finnish insect and spider species through the Barcode of Life Data system (BOLD) identification engine. Of these, 91% were correctly assigned to a unique species when compared to the new reference library alone, 85% were correctly identified when compared to BOLD with the new material included, and 75% with the new material excluded. To capitalize on this resource, we used the new reference material to train a probabilistic taxonomic assignment tool, FinPROTAX, scoring high success. For the full-length barcode region, the accuracy of taxonomic assignments at the level of classes, orders, families, subfamilies, tribes, genera, and species reached 99.9%, 99.9%, 99.8%, 99.7%, 99.4%, 96.8%, and 88.5%, respectively. The FinBOL arthropod reference library and FinPROTAX are available through the Finnish Biodiversity Information Facility (www.laji.fi) at https://laji.fi/en/theme/protax. Overall, the FinBOL investment represents a massive capacity-transfer from the taxonomic community of Finland to all sectors of society.

KEYWORDS
COI, DNA barcodes, probabilistic taxonomic assignment, PROTAX, reference library, species identification

1 | INTRODUCTION

Over the past decade, DNA-based identification has been adopted as a key tool for characterizing biological specimens (Hebert et al., 2016; Hebert, Ratnasingham, et al., 2016). To compare species composition among sites, to describe community organization, or to access previous knowledge related to the taxa encountered, specimens must first be identified. A quick and efficient approach is to cluster specimens into molecular operational taxonomic units or MOTUs (Blaxter et al., 2005). Indeed, the clustering of sequences combined with an interim taxonomy enables efficient characterization of biodiversity (Smith et al., 2013) and of species interactions (Clare et al., 2019). Yet, full realization of the value of such data relies on connecting as many MOTUs as possible to Linnaean taxonomy, because this makes it possible to connect species detected in DNA-based surveys to prior biological knowledge. Thus, the most efficient avenue for combining molecular data with taxon-specific knowledge involves populating reference databases with DNA barcodes annotated with Linnean taxonomy (Hebert et al., 2003). By definition, such progress can only be achieved through the active involvement of taxonomists.

Once populated, DNA barcode reference libraries can be used to establish the likely identity of a query sequence—and to partition the millions of reads from a high-throughput sequencing run to their likely source species. In such use cases, the reference sequence with the highest similarity (Altschul et al., 1990) is often assumed to represent the likeliest taxon, and its taxonomic tag becomes the relevant identification (BOLD Team, 2019). Importantly, any taxonomic placement made through this approach comes with uncertainty, since both the query and reference sequences may contain read errors, and because there is variation among sequences within a species. A further important source of uncertainty arises from the fact that reference sequence databases are incomplete, and they contain some incorrectly identified records (Pentinsaari et al., 2020). The best way to reduce uncertainty and improve performance involves extending species coverage and improving the quality of the reference databases (Meiklejohn et al., 2019; Pentinsaari et al., 2020; Wilson et al., 2011).

To arrive at comprehensive, reliable reference libraries, several nations and campaigns have constructed DNA barcode databases. Approaches range from barcoding all macroscopic species in an arctic region (Wirta et al., 2016) or a coral atoll (Andersen et al., 2019) to campaigns taking a taxonomic or geographic focus (e.g., Dincă et al., 2021; Miller et al., 2016; Zhou et al., 2016; for a summary, see http://www.ibol.org/phase1/about-us/campaigns/). Among the most ambitious initiatives to date are Fauna Bavarica, striving for coverage of all species in this German state (https://barcoding-zsm.de/bfb), and the intensive work in the Area de Conservacion Guanacaste, northwestern Costa Rica—with efforts to barcode all species in this nation (Janzen & Hallwachs, 2019; Miller et al., 2016). Other campaigns strive to generate comprehensive DNA barcode libraries for the biota of a country, including for example, Austria (ABOL, https://www.abol.ac.at/), Belgium (BeBOL, http://bebol.myspecies.info/), Germany (GBOL, https://bolgermany.de/home/en/german-barcode-of-life-2; see Morinière et al., 2019), Norway (NorBOL, http://www.norbol.org/), and

Switzerland (SwissBOL, http://www.swissbol.ch); see Figure 1. In each case, the combination of high species coverage with reliable taxonomic annotations is a key objective.

In this study, we consolidate knowledge from several sources to create a new tool that enables the taxonomic identification of more than 10,000 species, linking molecular samples with taxonomic collections and expertise. Specifically, we report the creation of a DNA barcode library for the arthropods of Finland, one built through a nationwide network of taxonomic experts. Ultimately, the Finnish Barcode of Life initiative (FinBOL, https://www.finbol.org/) will establish a DNA barcode reference library for all ~48,000 species of multicellular organisms that occur in Finland. The present study represents important progress toward this goal as it releases a reference data set for 11,275 arthropod species. We describe the approach employed to build this reference library, the current success rate, and the improvement in species identification resulting from its use. To capitalize on its development, we trained PROTAX, a probabilistic taxonomic assignment tool (Somervuo et al., 2016), to identify arthropod sequences from Finland—while accounting for gaps in knowledge and available reference material. Implemented as a web-based service, this new resource (FinPROTAX; https://laji.fi/en/theme/protax) allows both the accurate taxonomic placement of insects and the evaluation of the uncertainty associated with placements at each level in the taxonomic hierarchy.

2 | MATERIALS AND METHODS

2.1 | Approach

FinBOL was built using a crowdsourcing approach, as its progress reflects contributions by a network of about 150 Finnish taxonomists who contributed identified arthropod specimens for sequence analysis. This network involves both professional researchers and amateur naturalists. For details regarding the organization and activities of the network, see Appendix S1.

To maximize efficiency, FinBOL adopted a flexible strategy to obtain tissue samples of good quality for DNA barcoding. This flexibility has involved the utilization of both museum and private collections, the latter of which are many and of high quality in Finland, with different collectors focusing on different taxa. In return, FinBOL provided all participants with open access to the resultant data. Following this principle, FinBOL has not required that voucher specimens be deposited in public collections, but rather that they are maintained in known collections and only eventually—when the owner is deceased or discontinues collection—donated to museums.

Specimens obtained for DNA barcoding were mostly tissue sampled and photographed at the Zoological Museum of the University of Oulu and subsequently returned to their source public or private collection. To a lesser degree, some or all of these stages were carried out by the Finnish Museum of Natural History Luomus, by individual contributors, or, for photography, at the Centre for Biodiversity Genomics (CBG) at the University of Guelph, Ontario, Canada. Tissue samples (usually one or two legs, or a part of a leg, depending on the size of the specimen) were placed in 96-well microplates prefilled with ethanol and shipped to the CBG for DNA extraction and sequencing. For some minute species of Diptera, Coleoptera, or Acari, for which the regular tissue sampling approach was not feasible, plates containing whole specimens were assembled, and vouchers were recovered from the plates after DNA extraction using nondestructive methods.

DNA extraction followed a standard high-throughput protocol (Ivanova et al., 2006). A cocktail of the Folmer (Folmer et al., 1994) primers and LepF1 and LepR1 (Hebert et al., 2004) was then used to PCR amplify the target region of cytochrome *c* oxidase I in most specimens. The resultant amplicons were Sanger sequenced on an ABI 3730XL. Additionally, sequences of some taxa were produced in FinBOL participants' individual research labs, largely employing the protocols outlined above. The barcode sequences, as well as photographs and metadata for specimens, were uploaded to the Barcode of Life Data system (BOLD) (Ratnasingham & Hebert, 2007, 2013). Within the global database, all FinBOL projects are listed under the FinBOL campaign in the BOLD project list.

The current reference resource includes all FinBOL arthropod data uploaded and validated by October 31, 2019. All sequences with a minimum length of 500 bp and a validated taxonomic identification were compiled into a data set on BOLD (dx.doi.org/10.5883/DS-FINPRO). The full set of sequence data was downloaded as a time-stamped version and used to train the probabilistic taxonomic classifier PROTAX (see Section 2.3). As data will continue to accumulate well into the future, albeit at a slower rate, new sequences will be continuously uploaded to BOLD where they will remain connected to FinBOL through project identity (see above). All sequences that are not flagged as misidentified or contaminated are automatically included in the identification engine on BOLD (BOLD Team, 2019).

2.2 | Validating the taxonomic resolution achieved

To examine how much the national reference library improved the identification success of Finnish arthropods, we adopted a user's perspective. In brief, we examined the impact of the FinBOL arthropod material on species identifications generated by the BOLD Identification Engine—a web-based tool querying all sequences uploaded to BOLD from public and private projects to locate the closest match. Based on 1000 query sequences, we evaluated identification success when the query material was compared to BOLD under three scenarios: (1) BOLD without the new records, (2) BOLD with the new records added, and (3) BOLD restricted to the new records alone.

The BOLD ID Engine accepts sequences from the 5′ region of the mitochondrial COI gene and returns a list of closest matches to the query sequences. The user can choose between querying the full COI database or limiting the query to reference records with species-level identifications. For identification, BOLD uses the BLAST algorithm (Altschul et al., 1990) to identify single base indels

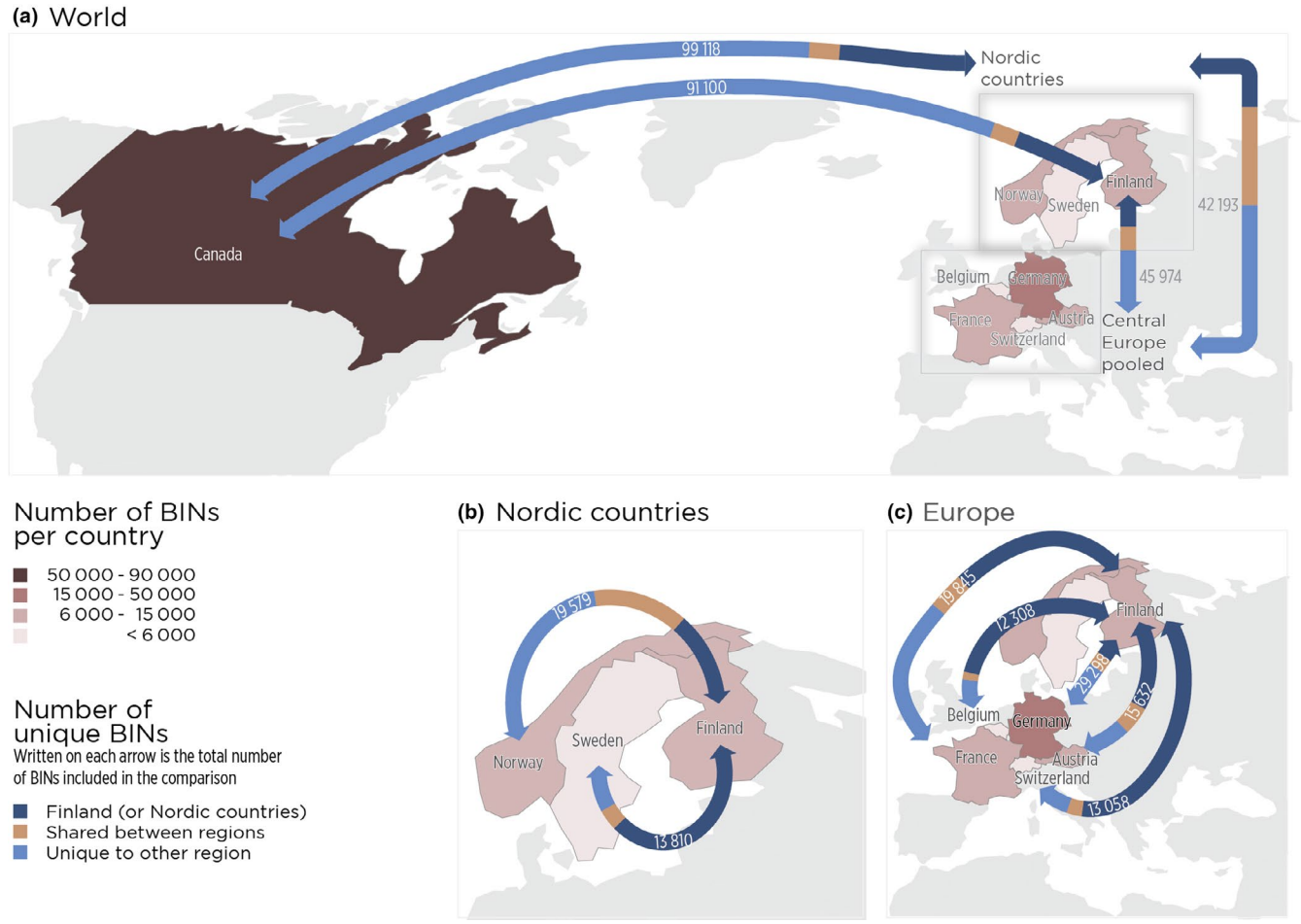

FIGURE 1 Complementarity in BIN composition of arthropod faunas between FinBOL and other regional DNA barcoding campaigns. Shown is the number of arthropod BINs unique to the Nordic countries (a) or FinBOL (a–c) with the arrows joining the two regions for which the comparison is made. Shades from pink to brown refer to the total number of BINs contributed to BOLD by each regional campaign. Regions in grey are not considered in this comparison (although some have contributed DNA barcodes to BOLD). Since all campaigns are in progress, the numbers for Finland refer to the current data release whereas the numbers for other areas refer to records on BOLD on 21 January 2021. In total, the FinBOL data release contained 13,777 BINs of which 1713 BINs had not been previously contributed to BOLD. For exact numbers, see Appendix S2

before aligning the protein translation through profile to a hidden Markov model of the COI protein (BOLD Team, 2019). As our reference library, we used the Species Level Barcode Records Database, that is, every COI sequence of >500 bp with species level identification uploaded on BOLD.

To assess whether and how much species-level taxonomic assignment improved with access to the FinBOL arthropod reference library, we selected a test set of 1000 Finnish arthropod species stratified by order. Species were chosen in rough proportion to national species-level diversity per order (Figure 2), with an important exception: to reduce the dominance of the four most diverse orders (Coleoptera, Lepidoptera, Diptera, and Hymenoptera) and increase representation of other taxa, we included only 150 species for each of these orders. The test set was assembled from the full FinBOL arthropod data set by first reducing the data set to only those species for which at least two sequences were obtained, and then randomly drawing the predetermined number of representatives from each order (with one sequence per species). This set of 1000 sequences was compared against the COI Species database on BOLD, that is, only including reference sequences with species level identifications, using the Batch ID Engine tool (accessed on March 8, 2021) with the default parameters: a minimum of 80% sequence similarity, and a minimum overlap of 300 bp between query and reference sequence.

The Batch ID Engine outputs a list of the top 100 closest matches to the query sequence in the database which fulfill these criteria, excluding self-match. If less than 100 matches for a given query sequence meet the criteria, the resulting list of matching sequence records is shorter. Each query sequence was assigned to the species with the best-matching reference sequence. Identification success (true or false) was evaluated assuming that the original identification of the query sequence was correct—a reasonable assumption since in each case, the identification had been made by the best available national expert (see section Approach, above).

Since the Batch ID Engine excludes comparisons to the query sequence itself, our restriction of the query material to species

with at least two reference sequences in the FinBOL material ensured the existence of at least one reference sequence in the FinBOL data set. Our three scenarios thus correspond to asking: how accurate an identification would we have achieved if querying the global database without the FinBOL arthropod records (scenario 1: comparison to BOLD with the new records excluded); how accurate an identification will we achieve now, when the FinBOL records are added (scenario 2: comparison BOLD with the new records included); and how accurate an identification will we achieve if we take the regional origin of the sequence into account, restricting our reference library to the national library alone (scenario 3: BOLD restricted to the new records only). For this purpose, we queried the Batch ID Engine for the taxonomic annotation associated with the best-matching sequence while excluding or including FinBOL sequences from the result list (corresponding to scenario 1 and 2, respectively), or excluding all non-FinBOL sequences from the list (for scenario 3). From the closest matches, we then determined the proportion of query sequences assigned to a single correct species ("Correct"), to several alternative species with the same sequence similarity of which at least one represented the correct one ("Several alternatives") or to the wrong species ("False").

2.3 | Training PROTAX to identify Finnish arthropods

To capitalize on the new resource, we trained PROTAX (Somervuo et al., 2016), a probabilistic taxonomic assignment tool, to identify arthropod sequences from Finland. PROTAX is a taxonomic classifier which establishes the probability that a query sequence can be assigned to a given taxon. In comparison to the identification engine used by BOLD, PROTAX has the advantage of recording how much the suggested identification can be trusted at each taxonomic level. PROTAX uses the sequences in the reference library to parameterize a statistical model of the probability with which a query sequence belongs to any particular taxonomic level (class, order, family, subfamily, tribe, genus, and species), or to a previously unknown taxon at the same taxonomic level. The latter probability should explicitly be interpreted as "a taxon not represented in the reference library", and thus explicitly accounts for current gaps in coverage.

2.3.1 | Taxonomy

Since PROTAX is based on hierarchical assignment to nested taxonomic levels, it builds on a fully resolved taxonomical hierarchy. However, it also accounts for the fact that there may be unknown taxa at each taxonomic level (see above) so the taxonomy represents both known and unknown insects of Finland. To train PROTAX, we used the names and taxonomic hierarchy of known taxa in the 2019 edition of the national checklist of Finnish species (FinBIF, 2020). Under each known genus, tribe, subfamily, family, order, class, and phylum in the taxonomy tree, there is also a branch corresponding to unknown taxa.

The taxonomic tree used to train PROTAX was constructed based on the full hierarchy and taxonomic names of 26,437 species. The root node of the tree represents the phylum Arthropoda, with a total of 48,801 nodes in the full tree covering the seven levels. Since the usage of taxonomic ranks varies greatly among taxa and taxonomists, the reference taxonomy used (FinBIF, 2020) contained some missing values for particular combinations of taxonomic levels and taxa. In those cases where a name was missing from the full taxonomic classification at a certain taxonomic rank, a dummy name was created. For example, the bird louse genus *Actornithophilus* belongs to the family Menoponidae in the order Phthiraptera, but subfamily and tribe ranks are not used in the reference checklist. Therefore, two dummy names were created to link this genus to its family. To create a single fully connected taxonomy tree, a total of 17,783 dummy names were introduced. For unknown taxa, that is, branches not included in the known taxonomy, an additional node was created under each internal tree node (see previous paragraph). These nodes allow for branches possibly missing from the set of known taxonomic names. For example, at the species level there were 33,335 nodes, of which 26,437 represented known species and the remaining 6898 nodes represented unknown species under 6898 known genera. Similarly, nodes representing unknown branches were added to the tree at all other levels of the taxonomy (but these nodes did not have further child nodes in our taxonomy tree).

To train the classifier in taxonomic assignment, we used 37,422 sequences from the FinBOL arthropod data set accompanying this study (Table 1). Of these sequences, 2798 sequences were only assigned to a genus or to an interim species, whereas 34,624 sequences were assigned to valid species. Out of 26,437 known species in the taxonomy, the data set used to train PROTAX included at least one reference sequence for 10,985 species. Out of 6898 known genera, the data set included at least one reference sequence for 3910 genera.

2.3.2 | Modelling approach

In PROTAX, classification starts at the root node of the taxonomic hierarchy, where a query sequence belongs with probability of one, and proceeds to leaf nodes passing through all ranks. Probability assignment from a parent node to its child nodes is achieved by means of a multinomial regression model. The parameters of the model are estimated using reference sequences to mimic query sequences coming from different parts of the taxonomy. A detailed description of PROTAX can be found in Somervuo et al. (2016), and a detailed description of the current implementation in Appendix S3. For the present purpose, the software has been rewritten in C to maximize its performance and is available in the github repository https://github.com/psomervuo/protaxA.

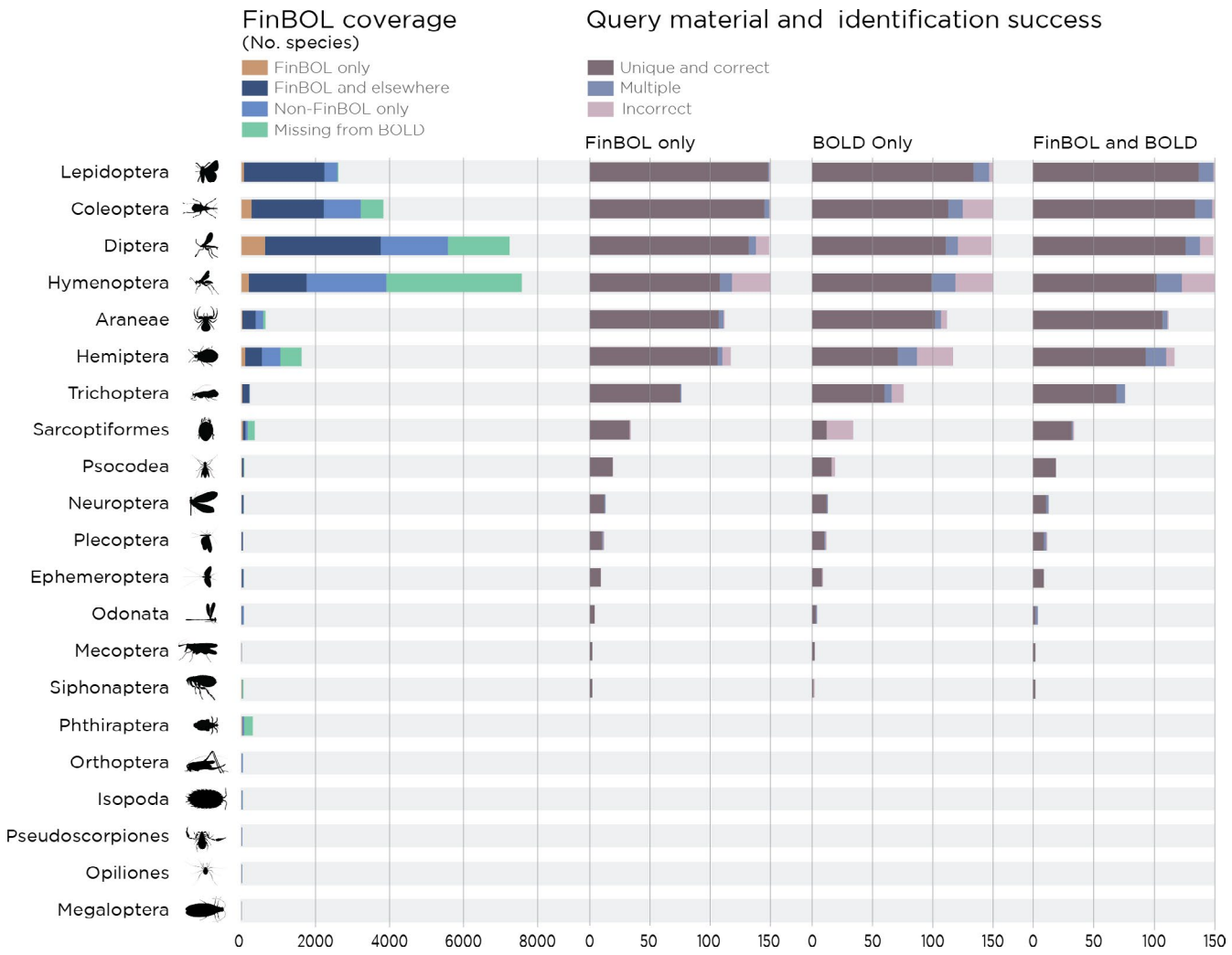

FIGURE 2 Taxonomic composition of the known Finnish arthropod fauna, its representation in reference libraries, and the identification success achieved by molecular tools. Shown on the left is the number of arthropod species per order in Finland, with the total length of each horizontal bar indicating total species richness and the sections within each bar showing the fraction of Finnish species for which DNA barcodes only occur in the FinBOL material (maroon); in FinBOL and in other BOLD material (dark blue); in BOLD but lacking from FinBOL (cyan); or completely missing from BOLD (green). The right-hand part of the figure identifies the improvement in identification success resulting from the FinBOL records. Shown from right to left is identification success under three scenarios: (1) BOLD without the FinBOL records, (2) BOLD with the FinBOL records added, and (3) identification by comparison to the FinBOL records alone. Sequences were assigned to species based on sequence similarity with reference sequences in BOLD using the BOLD ID Engine. Identification success was scored assuming that the original identification was correct. Sections within each horizontal bar show the proportion of query sequences assigned to a single correct species (column "Correct"), to several alternative species with the same sequence similarity, of which at least one represented the correct one (column "several alternatives") or to the wrong species ("False"). The composition of the overall fauna is taken from the Finnish national checklist of species (FinBIF, 2020). This checklist was also used to query the representation of Finnish species on BOLD. Due to possible differences between reference checklists used by different BOLD users when submitting data (e.g., in delimitation of genera), a single species may appear on BOLD under more than one name. As a result, our coverage counts for many orders are likely to be slight underestimates

To match typical data types, we constructed two versions of PROTAX, one for the full-length (658 bp) Folmer region (Folmer et al., 1994), adopted as the standard DNA barcode for animals (Hebert et al., 2003), and another for the Leray region (313 bp), as amplified by primers mlCOIintF (Leray et al., 2013) and jgHCO2198 (Geller et al., 2013). Parameter estimation was done using MCMC as explained in Somervuo et al. (2016). Model parameterization was done separately for each level of taxonomy as in Somervuo et al. (2017). For each of the seven levels of the taxonomy, 10,000 training sequences were generated from reference sequences and 2,000 iterations of MCMC were performed. The first half of the iterations was used for adapting the proposal distribution, whereas MAP estimates of parameters were selected from the second half of the iterations where the proposal distribution was fixed. The probabilistic taxonomic assignment tool parameterized for Finnish arthropods is henceforth referred to as FinPROTAX.

TABLE 1 Extent and coverage of the FinBOL arthropod reference library compared to overall arthropod species richness in Finland

Class	Order	Species richness	FinBOL			
			Individuals	Species	Unique species	BINs
Arachnida	Araneae	645	1553	354	13	373
Arachnida	Astigmata	14	0	0	0	0
Arachnida	Ixodida	3	0	0	0	0
Arachnida	Mesostigmata	437	0	0	0	0
Arachnida	Opiliones	17	2	2	0	2
Arachnida	Oribatida	352	302	105	27	146
Arachnida	Prostigmata	308	0	0	0	0
Arachnida	Pseudoscorpiones	18	3	2	0	2
Arachnida	Arachnida TOTAL	1794	1862	463	40	523
Entognatha	Collembola	244	0	0	0	0
Entognatha	Diplura	1	0	0	0	0
Entognatha	Protura	3	0	0	0	0
Entognatha	Entognatha TOTAL	248	0	0	0	0
Insecta	Archaeognatha	2	0	0	0	0
Insecta	Blattodea	8	4	0	0	1
Insecta	Coleoptera	3829	6921	2242	206	2350
Insecta	Dermaptera	3	3	1	0	1
Insecta	Diptera	7240	7724	3471	474	3545
Insecta	Ephemeroptera	56	93	25	1	30
Insecta	Hemiptera	1619	938	490	60	462
Insecta	Hymenoptera	7572	8627	1796	147	1927
Insecta	Lepidoptera	2616	10,382	2406	54	2488
Insecta	Mecoptera	7	11	6	0	6
Insecta	Megaloptera	5	1	1	0	1
Insecta	Neuroptera	61	161	42	1	42
Insecta	Odonata	62	31	17	0	16
Insecta	Orthoptera	34	1	1	0	1
Insecta	Phthiraptera	303	62	10	0	26
Insecta	Plecoptera	36	189	30	1	31
Insecta	Psocoptera	73	147	53	9	59
Insecta	Raphidioptera	3	0	0	0	0
Insecta	Siphonaptera	51	87	13	7	22
Insecta	Strepsiptera	9	0	0	0	0
Insecta	Thysanoptera	145	0	0	0	0
Insecta	Trichoptera	218	972	207	16	240
Insecta	Zygentoma	4	0	0	0	0
Insecta	Insecta TOTAL	23,956	36,380	10,811	976	11,248
Branchipoda	Anostraca	2	0	0	0	0
Branchipoda	Diplostraca	91	0	0	0	0
Branchipoda	Laevicaudata	1	0	0	0	0
Branchipoda	Notostraca	1	0	0	0	0
Branchipoda	Branchiopoda TOTAL	95	0	0	0	0
Malacostraca	Amphipoda	18	0	0	0	0
Malacostraca	Cumacea	1	0	0	0	0

TABLE 1 (Continued)

Class	Order	Species richness	FinBOL			
			Individuals	Species	Unique species	BINs
Malacostraca	Decapoda	9	0	0	0	0
Malacostraca	Isopoda	33	1	1	0	1
Malacostraca	Mysida	8	0	0	0	0
Malacostraca	Tanaidacea	1	0	0	0	0
Malacostraca	Malacostraca TOTAL	70	1	1	0	1
Maxillopoda	Arguloida	2	0	0	0	0
Maxillopoda	Calanoida	21	0	0	0	0
Maxillopoda	Cyclopoida	44	0	0	0	0
Maxillopoda	Harpacticoida	44	0	0	0	0
Maxillopoda	Poecilostomatoida	8	0	0	0	0
Maxillopoda	Porocephalida	1	0	0	0	0
Maxillopoda	Siphonostomatoida	16	0	0	0	0
Maxillopoda	Thoracica	1	0	0	0	0
Maxillopoda	Maxillopoda TOTAL	137	0	0	0	0
Ostracoda	Ostracoda TOTAL	74	0	0	0	0
Pauropoda	Pauropoda TOTAL	8	0	0	0	0
Symphyla	Symphyla TOTAL	6	0	0	0	0
Chilopoda	Geophilomorpha	10	0	0	0	0
Chilopoda	Lithobiomorpha	9	0	0	0	0
Chilopoda	Scolopendromorpha	2	0	0	0	0
Chilopoda	Chilopoda TOTAL	21	0	0	0	0
Diplopoda	Chordeumatida	1	0	0	0	0
Diplopoda	Julida	20	0	0	0	0
Diplopoda	Polydesmida	5	0	0	0	0
Diplopoda	Polyxenida	1	0	0	0	0
Diplopoda	Polyzoniida	1	0	0	0	0
Diplopoda	Diplopoda TOTAL	28	0	0	0	0

Notes: Species richness refers to the number of species reported in the FinBIF checklist of Finnish species 2019 (FinBIF, 2020). Under FinBOL, individuals and species refer to the number of barcoded individuals and species, respectively, in the FinBOL data set released with this study and used to train the PROTAX implementation for probabilistic taxonomic assignment. "Unique species" is the count of species only contributed to BOLD though FinBOL, that is, for which there would otherwise be no reference sequence. Three classes with one order in each (Pauropoda, Symphyla, Ostracoda) have each been compressed into a single line "TOTAL". For this reason, the table only includes 60 order-specific lines although 63 arthropod orders are known from Finland. For a visual representation of FinBOL coverage, see Figure 2.

2.4 | Validating FinPROTAX performance

To validate the performance of the parameterized PROTAX model, we used two approaches:

First, we used 10,000 sequences from the FinBOL arthropod data set as query sequences. Since the correct assignment was known to the species level for each of these sequences, we could validate the accuracy with which a query sequence was attributed to the correct taxon at each level in the taxonomic hierarchy (class, order, family, subfamily, tribe, genus and species). In this validation study, we calculated the probabilities from each query sequence against all taxa and took the taxon corresponding to the highest probability. During this process, the query sequence was removed from the reference sequences so the query sequence was not allowed to match itself. The assignment was deemed correct if the taxon with the highest probability matched the given taxonomic label of the sequence. In this way, we were able to compare how the best probability given by PROTAX corresponds to the correct classification and verify that the probabilities provided by PROTAX are unbiased (Somervuo et al., 2016). To speed the search, we excluded all taxa with negligible identification probabilities, using 0.01 as the threshold. If a parent node had a probability below the threshold, all child nodes of the parent were excluded from the further search. This test was run separately for the full barcoding region and the Leray region.

Second, we assessed the taxonomic confidence a user of FinPROTAX will achieve in identifying an environmental sample

of Finnish insects. For this purpose, we use an increasingly common type of highly diverse samples (Barsoum et al., 2019; deWaard et al., 2019; Lopez-Vaamonde et al., 2019): the catches from a Malaise trap (Geiger et al., 2016; Malaise, 1937; Townes, 1972). The trap in question was run for a full summer in 2012 (20 weeks from 16 May to 3 October) in Kiiminki near Oulu, Finland (coordinates 65.148 N, 25.838 E). All insects in the collections were individually Sanger sequenced using the methods described in Appendix S4. Of the resulting 6486 sequences, we ascertained the fraction that could be assigned to a given taxonomic level. Realizing that different researchers will be satisfied with different levels of confidence, we carried out this analysis at two probabilities: 0.9 and 0.5. For interpreting these cutoffs as "reliable" (0.9) and "plausible" (0.5), we refer the reader to section Training PROTAX to identify Finnish arthropods and to (Somervuo et al., 2016, 2017), noting explicitly that these probabilities naturally build on sequence similarity, but otherwise have nothing in common with a simplistic cutoff of, say, 98% or 99% sequence similarity (e.g., Clare et al., 2019).

3 | RESULTS

3.1 | A COI barcode library of Finnish arthropods

This study makes available 37,422 reference sequences, of which 92.5% (34,624) are assigned to a known species. From the 26,437 arthropod and 23,956 insect species known from Finland, the FinBOL library providers coverage for 11,275 and 10,811 species respectively, with coverage reaching 92% for the 2616 species of Lepidoptera and 94% for the 218 species of Trichoptera (Table 1). Current coverage is strongly biased towards insects and spiders with low to no coverage for other arthropod groups. Mites in the order Oribatida are an exception as their higher representation reflects a targeted national campaign and a dedicated taxonomist (R. Elo).

In terms of taxonomic coverage, the national barcode library for Finland is complementary to the global database BOLD and to other national barcoding campaigns. In total, the current 11,275 species correspond to 13,777 BINs of which 12.4% (1713) are new to BOLD. *Mutatis mutandis*, of 608,360 arthropod BINs currently represented on BOLD (accessed 26 March 2021), 2% occur in FinBOL. The correspondence between BINs and species in the Finnish fauna has been examined in previous, taxon-specific studies (see Discussion for detailed references).

At a large spatial scale, the BIN content of the Finnish barcode library emerges as strongly complementary to that of more distant geographic regions such as North America and Central Europe (Figure 1a). Viewed at a regional scale, it is complementary to the modest DNA barcoding efforts in Sweden and the extensive campaigns in Norway (Figure 1b). Considering national barcoding campaigns in Europe, the Finnish effort is substantial and complementary in terms of BIN coverage (Figure 1b; for exact numbers see Appendix S2). Needless to say, the fact that one in eight BINs contributed to BOLD is new also indicates that the rest were earlier sequenced somewhere else. Thus, BIN overlap was also substantial (Figure 1).

3.2 | Taxonomic resolution achieved

The new records contributed by FinBOL substantially improve upon the accuracy achieved by the global identification resources. From our query material of 1,000 insect and spider species, 73% of insects and 91% of spiders were correctly assigned a single, unequivocal best match by BOLD when the Finnish material was removed (Figure 2) versus 84% for insects and 96% for spiders once it was added. This was mainly due to fewer false assignments with a smaller reduction in the fraction of taxa yielding multiple equally-well matching identifications (Figure 2). The highest accuracy was achieved when comparison was restricted to the national reference library alone (Figure 2). Identification success varied substantially among orders, but reached 100% in several well-represented orders (Figure 2).

The taxonomic coverage of the FinBOL arthropod reference data (Figure 2, left-hand part) allowed accurate training of FinPROTAX. Because certain classes were absent from the training set (Figure 2, right-hand parts; Table 1), the training of FinPROTAX was restricted to spiders (order Araneae) and insects (class Insecta) only. For these taxa, the taxonomic classifier achieved high accuracy in assigning query sequences to the correct taxon at all taxonomic levels. For the 10,000 query sequences of known taxonomic affinity, the probabilities proved unbiased sensu Somervuo et al. (2016). In other words, if PROTAX assigns a probability 0.9 for a query sequence, then for any large number of sequences, one in ten will be incorrectly classified while nine of ten will be correct. The same applies for all probabilities, that is, if PROTAX gives a probability p, then $100p$% of such sequences are classified correctly and $100(1 - p)$% of the sequences are incorrect. For the full-length Folmer region, the accuracy of taxonomic assignment (i.e. the proportion of units within the respective taxonomic group assigned to the correct taxon) was 99.9% at the level of both classes and orders, 99.8% at the level of families, 99.7% at the level of subfamilies, 99.4% at the level of tribes, 96.8% at the level of genera, and 88.5% at the level of species. For the Leray region, accuracies were very similar (99.9% for classes and orders, 99.6% for families, 99.4% for subfamilies, 99.1% for tribes, 96.6% for genera, and 87.8% for species).

3.3 | With what accuracy can we identify an environmental sample?

The accuracy achieved by FinPROTAX is further demonstrated by its classification of an independent data set from a Malaise trap. Among the 6486 sequences, some were assigned to known taxa with a high probability whereas others were not (Figure 3, for an interactive graph see Appendix S4). As expected, the probability of correct

assignment was greatest for higher levels in the taxonomic hierarchy, for which more training data was available (Table 2; Appendix S4). Interestingly, differences in assignment success between the two cutoff levels (0.9, 0.5) were surprisingly small: if a sequence was attributed to an unknown branch at a lower level, it was likewise consistently attributed to an unknown branch at a higher level (Table 2).

Among Insects, a substantial proportion of sequences were assigned to the category of "unknown order" (Figure 3). This category included sequences for which the highest probability included taxa not represented in the reference sequences. They included both taxa not represented in the current taxonomy and known taxa lacking reference sequences. Closer investigation revealed that most of these sequences were derived from two species, *Lepidocyrtus lanuginosus* (Class Entognatha, Order Collembola, Family Entomobryidae) and *Euceraphis punctipennis* (Class Insecta, Order Hemiptera, Family Aphididae). Both species occur in Finland (FinBIF, 2020), but the current set of reference sequences lacks representatives of them even at the class (for *Lepidocyrtus lanuginosus*; Table 1) or family (for *Euceraphis punctipennis*) level.

Among orders, the level of confidence in identifications varied substantially among taxa. For Lepidoptera, sequences were relatively evenly and reliably assigned to a species, reflecting its high coverage in the reference database (Figure 1; Table 2). For Diptera, out of 1,918 sequences that were classified to the order level with a probability greater than 0.90, 915 sequences (48%) were assigned to a known species with a probability exceeding >0.9. For Hymenoptera, the corresponding numbers were 1,088 and 130 (12%) much lower than those for Lepidoptera (123/150 = 82%), for Oribatida (106/112 = 95%), and for Araneae (60/71 = 85%). Among mites in Oribatida, most sequences were classified into a single species, *Diapterobates humeralis*. By contrast, a large proportion of sequences for Hymenoptera were identified with substantial uncertainty (see large grey circle among Hymenoptera in Figure 3).

4 | DISCUSSION

Building on a decade of work, the Finnish Barcode of Life initiative has delivered a unique reference library for specimen identification and a parameterized tool for probabilistic taxonomic assignment. The main outcome is a massive transfer of identification capacity from the taxonomic community to society at large. In the process, the project provided five valuable lessons about how national resources for biodiversity research may be built. The following text discusses each of these aspects.

4.1 | A national reference library for arthropods: Extent and coverage

Of the 26,437 arthropod species known from Finland, 23,956 of which are insects (FinBIF, 2020), the current FinBOL data release provides reference sequences for 11,275 and 10,811 species, respectively—i.e. 43% and 45% of the known national fauna. In well-studied groups, barcode coverage is high, with 92% of 2616 species of Lepidoptera and 94% vs 218 known species of Trichoptera. However, this relatively high coverage also extends to many "difficult", globally poorly known groups such as selected families of Diptera and Hymenoptera (Figure 2).

Despite the high coverage, the proportion of arthropod taxa which were contributed *uniquely* through FinBOL to the global database BOLD (Ratnasingham & Hebert, 2007, 2013) averaged just 7%, but ranged from 0 to 54% of the species per order (Table 1). The current figures are slightly inflated because a few species collected outside Finland were included in the data set. The relatively low uniqueness of the Finnish fauna reflects the recent deglaciation of the target area along with general rules of biogeography. The latest glaciation in Finland, the Weichselian, peaked 22,000 BP and although deglaciation began by 13,000 BP, the northern parts of Finland only became ice-free about 10,500 BP (Donner, 1995). As a result, almost all current species have colonized since then, and very few species are endemic. Species range size also tends to increase with latitude—a pattern known as Rapoport's rule (Ruggiero & Werenkraut, 2007; Stevens, 1996). Combined with a general decline in species richness with latitude (e.g., Hillebrand, 2004; Schemske & Mittelbach, 2017), this rule leads to fewer unique species per unit area with latitude. For a high-latitude country such as Finland (59°48′N to 70°05′N), considerable faunal overlap with neighbouring countries is expected.

Notwithstanding this overlap, the FinBOL arthropod initiative has contributed thousands of unique BINs beyond those submitted by any other region. This partly reflects differences in regional faunas, and partly sampling effects. Given the low coverage of DNA barcode libraries in Sweden (Figure 1 and Hovmöller et al., 2017) and Russia (data scarce and not shown), FinBOL has generated a valuable regional resource.

Against this general setting, the building of the national reference library has helped to clarify which species are truly shared with other regions. For example, a large-scale comparison of Austrian and Finnish Lepidoptera demonstrated that several arctic–alpine species thought to be shared between high latitudes and high elevations are actually different species (Huemer et al., 2014; Mutanen, Hausmann, et al., 2012). A similar comparison at an intercontinental level revealed many overlooked species shared between Palearctic and Nearctic regions (Landry et al., 2013; Pentinsaari et al., 2020). Comparisons of national barcode libraries regularly reveal cases of deep sequence divergence requiring detailed taxonomic study, and such work often leads to the description of new species. One recent example is the split of the charismatic moth *Pyralis regalis* (Denis & Schiffermüller, 1775, into two distinct but morphologically similar species (Wikström et al., 2020). Such results make clear the need to barcode representatives of presumptively widespread species across their range. While a reference sequence for a given species from any point in its range is valuable, a DNA barcode reference library based on restricted geographic sampling will only contain a

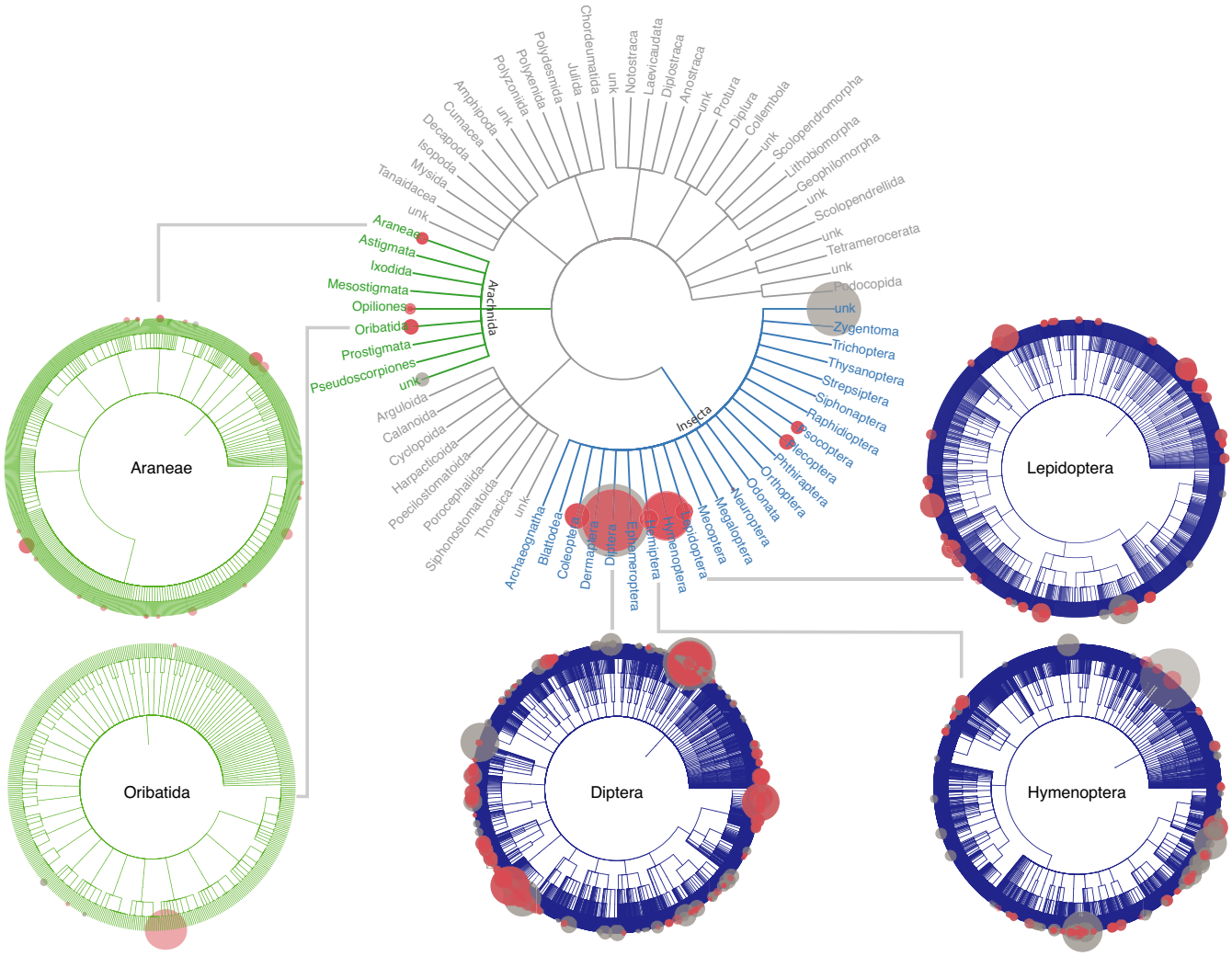

FIGURE 3 Resolution achieved in the taxonomic assignment of environmental data. Barcode sequences >500 bp were recovered from 88% (6,486/7,414) of the arthropods in an annual Malaise trap sample. For these organisms, we show the proportions of taxa within the phylum Arthropoda assigned to a given taxon. Each sequence was assigned to the taxon with the highest probability (excluding sequences where the highest probability was <0.1). Grey coloration indicates all sequences while red indicates the number of sequences assigned with a probability exceeding 0.9 (the sizes of the different diagrams are not internally comparable). The central diagram shows the classification to an order level. All samples were classified into two classes, Insecta and Arachnida. Five diagrams surrounding the central one show the classifications from the order to the species level. Araneae and Oribatida were the two largest orders for Arachnida while Diptera and Hymenoptera were the two largest orders for Insecta. To visualize the full contents of the sample, we provide an interactive Krona wheel (Ondov et al., 2011) in Appendix S4

fraction of the variation within most species (Bergsten et al., 2012; Huemer et al., 2014; Mutanen, Hausmann, et al., 2012), weakening its capacity to generate reliable taxonomic assignments on a global scale. For this reason, it is important to construct regionally comprehensive reference libraries of widespread species—both from a national and global perspective.

4.2 | Taxonomic resolution achieved

When added to the identification engine in BOLD, the barcode records generated by FinBOL substantially improved identification success. This improvement occurred against a background of already high success (see Figure 2)—even with all Finnish records excluded, a full 74% of the species in our query set were assigned to a single, correct taxon when queried through the BOLD identification engine. This attests to the power of the global barcoding effort, which has populated the global reference library with extensive, taxonomically annotated reference sequences—including massive European material (Figure 1).

For every major order (barring Hymenoptera; Figure 2), the sequences contributed by FinBOL produced a significant increase in identification success (Figure 2). One aspect of this success involved a general reduction in the number of false assignments, that is, cases where the best-matching sequence was erroneously annotated. Although some of these mismatches may reflect disagreement among taxonomists regarding the correct species name, such cases probably represent a minority of the "false" species assignments. In

TABLE 2 Taxonomic resolution achieved for a full-season sample of arthropods from a Malaise trap

	Cutoff 0.9		Cutoff 0.5	
Level	Known%	Unknown%	Known%	Unknown%
Class	88.1	0.1	99.7	0.2
Order	62.1	0.0	70.0	19.6
Family	54.9	2.1	60.2	5.6
Subfamily	47.8	2.0	53.2	5.3
Tribe	46.9	0.0	51.9	0.2
Genus	41.4	0.1	44.9	4.2
Species	29.4	3.1	34.1	7.6

Notes: Shown is the proportion of taxa assigned to a given taxonomic level with a probability exceeding a particular cutoff value, either 0.9 (left-hand columns) or 0.5 (right-hand columns). "known%" refers to branches with a reference sequence in the training set while "unknown%" are branches where no reference sequences available. For details on the specific data set and for an interactive visualization of the full contents of the sample, see Appendix S4.

other cases, BOLD users may have employed different checklists when submitting records, resulting in the same species appearing under different names in the database. For example, there are cases where the same species is named differently in the Old and New World, cases that await taxonomic revision and the declaration of synonyms (e.g., van Nieukerken et al., 2016). More importantly, since BOLD is not only a reference database, but also a workbench for analysing and curating DNA barcode data, some misidentified, contaminated, chimeric or otherwise erroneous sequences make their way into the database and may not be immediately excluded (see Pentinsaari et al., 2020). In addition, the species-level resolution of the COI barcode region is not always perfect as closely related species may share identical haplotypes or form mixed sequence clusters (Hausmann et al., 2013; Huemer et al., 2014; Prous et al., 2016). In such cases, multilocus approaches sometimes improve identification success (Meiklejohn et al., 2019), and assignment based on whole genomes may improve success even further (e.g., Ji et al., 2020). Yet, for establishing a national reference library of the current size, coverage, and curation since 2010, under then prevailing analytical costs, no realistic alternatives to the current single-locus approach were available. Here, the improvement in correct taxonomic assignment shows the value of a highly curated reference database, as achieved by the large network of skilled taxonomists contributing expert-identified material to FinBOL.

Overall, the improvement in identification success enabled by the FinBOL records seemed roughly proportional to species diversity in the target group (by affecting the number of choices) and by the representation of species in FinBOL versus the rest of BOLD (by affecting the added precision brought by the FinBOL effort). Clearly, the highest identification success is attained when the query sequences are matched to the national database alone (Figure 2). It should be noted, however, that the exact comparison performed here is restricted to taxa represented by at least two sequences in the FinBOL data—otherwise, we could not compare the match of query sequences to FinBOL versus non-FinBOL material. For some 15,000 arthropod species, of which 13,000 are insects, no reference material exists in FinBOL, and the lack of coverage is as high as 100% in some arthropod orders (Table 1; Figure 2). Thus, for a random sequence generated de novo from an unknown query sample, identification success varies substantially depending on the order (compare Figure 2). While the national checklist of arthropods is itself incomplete, it provides a vivid illustration of the challenges caused by the lack of reference sequences from the national barcoding library. For about one-half of all Finnish arthropod and insect species, no reference exists in FinBOL, and for about one-third none exist elsewhere in BOLD (Figure 2). Thus, if we assume that the reference sequence providing the best match to the query sequence will equal the correct identification, we will be off in a substantial proportion of cases. To control for this strong bias, our PROTAX implementation, FinPROTAX, quantifies the likelihood that the query sequence represents a species currently missing from the reference library.

4.3 | A new tool for taxonomic assignment

Some species will always be missing even from comprehensive DNA barcode reference libraries—simply because they are excessively rare or hard to obtain. Such gaps are caused either by incomplete sampling and/or by the fact that communities at all spatial scales show changes over time (e.g., Antão et al., 2020). Gaps in the reference library can be forgotten when taxonomic assignment is based on highest sequence similarity alone, even though these gaps may have a critical impact on the identifications achieved (see Somervuo et al., 2017). For FinBOL, the fact that nearly half of the national arthropod fauna has now been covered also implies that more than half awaits sequencing. The PROTAX implementation based on the Finnish reference library, FinPROTAX, allows researchers to account for such influences, and provides intuitive measures of uncertainty to help them evaluate the reliability of taxonomic assignments. Implemented as a web-based service (https://laji.fi/en/theme/protax), this new resource allows the accurate taxonomic placement of insect samples and the evaluation of the uncertainty associated with their placement at each level in the taxonomic hierarchy.

When validated by query sequences from the national reference library, the accuracy of FinPROTAX was high with 88.5% of test sequences being assigned to the correct species as the most likely match. The same unbiased result was achieved both for the full-length Folmer region and the Leray region—despite the latter being only half the length of the former (313 bp vs. 658 bp), and thereby including less nucleotide variation. Yet, beyond taxonomic assignment—that is, likely names with which to label the species—FinPROTAX also provides a measure of uncertainty—that is, of the probability with which this label is correct. The importance of this added consideration was highlighted by our application of FinPROTAX to a diverse sample of arthropods collected by a Malaise trap.

Notably, no algorithm can reliably attribute a sequence to a named taxon for which it has seen no training data. Thus, when PROTAX reports "unknown" taxa with high probability, it does not necessarily mean that the sequence originates from a previously unknown taxon not included in the taxonomy. Instead, it means that no good match was present among the existing reference sequences. To understand the implications of this outcome, it is important to consider what PROTAX does; it converts sequence distances into taxon probabilities. Hence, the quality of the sequences (both query and reference sequences) is key to accurate taxon assignments, so uncertainty can be due either to sequencing errors and/or to taxa not being absent from the reference sequences. An extreme case occurs when the taxon is included in the taxonomy but lacks any reference sequences. When PROTAX reports unknown taxa, it includes in that category also known taxa from which there are no reference sequences available (Somervuo et al., 2016).

This consideration is further illustrated by the large uncertainty associated with specific taxa in the Malaise trap material (Figure 3). Here, sequences from classes, orders, or species missing from the FinBOL material were consistently assigned to the category of "unknown" taxa—just as they logically should, since they are by definition unknown to the taxonomic classifier. In this context, we note that we only trained the classifier on material containing insects (class Insecta) and spiders (class Arachnida, order Araneae). This is because insects and spiders have been more popular among contributing taxonomists than any other arthropods (Table 1)—for which reason they are also the ones for which classification needs will most often arise. By definition, this implies that the current FinPROTAX classifier has been explicitly optimized for these two taxonomic groups, whereas it will face challenges in classifying other sequences (i.e. those from taxonomic groups never shown to FinPROTAX). In practice, arthropods outside of Insecta will either be attributed to the "unknown" category under the root node, or to "unknown Insecta".

Within classes, there was much variation in the probability of assignment for individual taxa. Between the levels of class and order, the differences in uncertainty were relatively small (Table 2). This outcome can probably be attributed to the fact that at this level, the barcode region lacks much phylogenetic resolution, since it codes for a crucial protein and all viable mutations have already been tried multiple times during evolution (i.e., variation is saturated; Pentinsaari et al., 2016). Within orders, the uncertainty associated with finer classification within families, genera and species was higher for diverse taxa with lower coverage in the FinBOL database, such as Diptera: Cecidomyiidae, Diptera: Sciaridae and Hymenoptera: Chalcidoidea (see Figure 3, Appendix S4). Again, this is a logical outcome: the less knowledge we have of what taxa to choose from and of the molecular variation within versus between taxa, the more difficult it is to attribute a DNA sequence to a specific taxon.

Importantly, the current taxonomic resolution achieved by FinPROTAX is based on the reference data alone, and neglects additional considerations. An obvious source of uncertainty not modelled by PROTAX involves errors resulting from PCR and sequencing. Therefore, when using PROTAX with sequences coming from error-prone high-throughput sequencing platforms, third-party software such as DADA2 (Callahan et al., 2016) should be used to reduce the incidence of errors in reads. The presence of sequencing errors does not prevent the use of FinPROTAX, but the errors will result in a higher uncertainty, i.e. less specific taxon assignments.

Another complication to taxonomic assignment is barcode sharing, i.e., cases where two valid species share the same sequence for the gene region examined. Such cases will naturally increase uncertainty as there is no way to tell such species apart. In ambiguous cases (see examples in Appendix S4), any additional data will be worth evaluating—including cases where extant ecological knowledge may extend the insights from molecular data. Where a resolved species-level taxonomy is missing for key groups, applications can be strengthened by mapping the geographic distributions of sequence-based species proxies (see Pentinsaari et al., 2020). As a future improvement to PROTAX, we propose to add priors informed by the spatial, temporal, and ecological context of the sample. As a simple solution, the user may a priori weigh their belief that a given species may occur in the sample based on, for example, digital maps of the distribution of the target taxa, their host plants, and habitats. If implemented as a simple dichotomy (e.g., probably vs. almost impossible species), this can be done for even a large number of species with reasonable effort.

4.4 | Lessons learnt

Obtaining funding for a national project like FinBOL was difficult. Importantly, the building of a national genetic resource was hard to frame as a primary research project, and was therefore initially declined by national funding agencies. Thus, the initial funding strategy for FinBOL relied on private foundations and infrastructure initiatives (Appendix S1). Despite initial skepticism from national funding agencies, FinBOL has delivered primary research results of the very type traditionally supported by them. Deliverables to date include insights into the protein structure of the barcode region (Pentinsaari et al., 2016), species interactions (Kaartinen et al., 2010; Mutanen, Ovaskainen, et al., 2020; Nyman et al., 2015; Rytkönen et al., 2019; Vesterinen et al., 2013, 2016), taxonomy (Boonstra et al., 2018; Haarto & Ståhls, 2014; Hausmann et al., 2013; Huemer et al., 2020; Huemer & Mutanen, 2015; Ivanov et al., 2018; Kirichenko et al., 2016; Kozlov et al., 2017; Landry et al., 2013; Landvik et al., 2013;

Lee et al., 2020; Liston et al., 2019; Mutanen, Aarvik, Huemer, et al., 2012; Mutanen et al., 2012, 2013, 2015, 2016, 2020; Nieminen et al., 2018; Pentinsaari et al., 2014; Pentinsaari et al., 2014; Pilipenko et al., 2012; Pohjoismäki & Haarto, 2015; Pohjoismäki et al., 2016; Prous et al., 2016, 2020; Pykälä & Myllys, 2016; Salmela et al., 2014; Sihvonen et al., 2020; Ståhls et al., 2015; Tabell et al., 2019; van Nieukerken et al., 2012; Wikström et al., 2020; Wilson et al., 2011), phylogenetics (Heikkilä et al., 2014; Kaila et al., 2020; Karsholt et al., 2013), faunistics (Korpelainen & Pietiläinen, 2017; Paukkunen & Kozlov, 2015), biogeography (Huemer et al., 2014, 2018; Mutanen, Hausmann, et al., 2012; Salmela, 2012), methodology (Kekkonen et al., 2015; Korpelainen & Pietiläinen, 2019; Korpelainen et al., 2016; Lee et al., 2018; Pentinsaari et al., 2017), environmental change (Keret et al., 2020) and life-history evolution (Kivelä et al., 2020).

Assessing our progress, we believe the FinBOL initiative has catalyzed national biodiversity research on a broad scale. A uniting factor for this scientific community is a need for species-level understanding to examine the emergent features of overall diversity. In this context, the project has overcome key aspects of the taxonomic impediment. It has truly been an investment in the construction of a national infrastructure for biodiversity science, adding an accurate and cost effective tool to Finnish science—just as one might regard investment in a telescope or particle accelerator (Hebert, Hollingsworth, et al., 2016; Hebert, Ratnasingham, et al., 2016). Clearly, the impact of this investment has extended beyond national borders. In fact, most of the taxon-specific initiatives and participants have engaged in international projects, where data generated by the Finnish initiative has played a key role in supporting investigations of continental or global patterns (e.g., Hausmann et al., 2013; Huemer et al., 2014, 2018; Landry et al., 2013; Mutanen et al., 2013, 2016; Pohjoismäki et al., 2016). Overall, the FinBOL initiative reinforces the evidence (Miller et al., 2016) for the utility of DNA barcodes in supporting biodiversity inventories (Wirta et al., 2015, 2016) from facilitating the identification of freshly collected specimens (Figures 2 and 3; Appendix S4), to linking the fresh specimens, museum collections and past taxon-specific knowledge, in the process flushing out cryptic species (Kaartinen et al., 2010; Landvik et al., 2013; Miraldo et al., 2014) and revealing unexpected synonyms (Haarto & Ståhls, 2014; Hausmann et al., 2013; Landry et al., 2013; Mutanen, Hausmann, et al., 2012; Prous et al., 2016; Sihvonen et al., 2020).

When FinBOL was initiated, there was antagonism between "traditional" taxonomists and the DNA barcode initiative (e.g., Ebach & Holdrege, 2005; Will et al., 2005), followed by calls for unification (Hebert & Gregory, 2005; Padial & De La Riva, 2007). During its decade-long activity, FinBOL has managed to close this gap while the international rift has more or less mended (e.g., Hebert, Hollingsworth, et al., 2016; Hebert, Ratnasingham, et al., 2016; Meierotto et al., 2019; Mutanen et al., 2013; Padial & De La Riva, 2007; Ratnasingham & Hebert, 2013). Among specific factors allowing the current success of FinBOL, we identify the communal and open approach adopted. To obtain taxonomically annotated tissue samples of good quality, we benefitted from both museum and private collections. In return for these contributions, FinBOL provided all participants with open access to the resultant data. Following this principle of openness, FinBOL has not required that voucher specimens be deposited in public collections, but rather that they are maintained in known collections and only eventually donated to museums. By this approach, the vouchers can be accessed if needed without the disruption to private collections.

During its activity, FinBOL has also actively encouraged synergy among taxonomists, ecologists, and other domains of biodiversity science. As a vivid illustration of this claim, the Malaise trap material used as our test case for FinPROTAX represented an insurmountable identification challenge to taxonomists. It contains close to 6,500 individuals, most representing common species, but all call for tedious sorting and many call for added preparations and dissections to confirm their identification. No single taxonomist can morphologically identify all the species which we detected by molecular analysis. By identifying these specimens by molecular means, we shift the efforts of expert taxonomists from routine sorting of bulk samples to the study of focal specimens. This we can do by pointing them directly to the most interesting individuals identified by DNA barcodes—be they rare, ecologically interesting, previously unknown, or lacking a previous reference sequence.

4.5 | Future directions

The FINBOL reference library can already deliver reliable taxonomic assignments for many taxa, but much remains to be done. In evaluating the finer details of the taxonomic assignments possible for our environmental sample, we note that the current evaluation applies strictly to a specific sample of a specific type. While highly diverse in species composition, any single sample naturally comes with its particular features. For example, the sampling site chosen for analysis will have an effect since some regions in Finland have been comprehensively sampled while other areas lacking museums or universities have not. The sampling technique employed will also affect the outcome. For example, the insects dominating a Malaise trap catch are better represented in the current reference library than are soil arthropods. To gain both regional and ecological coverage, efforts will be directed toward expanding the FinBOL arthropod library and the FinPROTAX tool based on new records, thereby improving its performance across all types of samples. Current plans call for 50% coverage for the estimated 48,000 multicellular species in Finland (Hyvärinen et al. 2019) being achieved in BOLD by 2022 and 90% by 2030.

Two methodological advances are aiding FinBOL's progress. One is the move to high-throughput sequencing platforms, such as SEQUEL (Hebert et al., 2018), which reduce analytical costs and increase throughput. A second advance involves the improved ability to recover DNA barcodes from old specimens (Prosser et al., 2016), enabling the use of museum collections to fill gaps in barcode coverage for rare species. FinBOL is currently sequencing large numbers of museum specimens representing rare taxa, including type material, to fill gaps. In the same way as type specimens were introduced to anchor the species name unequivocally to morphology, this

approach will allow us to add a sequence to the species description (Hausmann et al., 2016; Miraldo et al., 2014; Prosser et al., 2016).

While technological progress is accelerating the acquisition of sequence information, the ranks of taxonomic specialists are thinning. Many of the older members of Finland's taxonomic community feel that their contribution to the current reference library forms a lasting contribution to science. The need to capture such knowledge is essential because there are, for example, no young Finnish taxonomists who can critically identify species in many key groups of arthropods (e.g., aphids, chewing lice, chalcid wasps, gall midges, most mite lineages). Hence, the annotated barcode records assembled by FinBOL participants represent a tremendous intergenerational transfer of taxonomic knowledge. The ultimate aim of FinBOL is to facilitate the rapid, accurate identification of specimens regardless of life stage, sample size, and quality. Until now, this has demanded access to experts, but the accurate identification of voucher organisms is a demanding, time intensive process. However, the time contributed by current taxonomists in identifying and contributing voucher specimens represents a great gift to future generations who will benefit from their expertise when they are no longer able to process new material. Thus, the current contribution offers a major capacity donation from the taxonomist community to science. The many taxonomists among us rejoice at this opportunity to contribute while the many molecular biologists, ecologists, and arthropod enthusiasts feel gratitude for this contribution.

ACKNOWLEDGEMENTS

FinBOL could not have achieved success without the contributions from many people with species expertise in diverse arthropod groups. While most significant contributors of FinBOL are co-authors on this publication, many others with smaller yet important contributions are not. They helped by providing specimens for tissue sampling often from their private collections and by identifying samples, but also in multiple other ways. We are most grateful to all these persons. Also, FinBOL's work has been supported by a number of technicians, research assistants and students who have contributed in tissue sampling, photography and databasing of barcoded specimens. We are especially grateful to Piia Partanen and Riikka Jarkko for their major input with this regards. Staff at the Centre for Biodiversity Genomics (Guelph, Canada) have provided generous and continuous help in handling and analysing samples, curating data on BOLD, and helping to publish these records, for which we feel much indebted. The goals of the FinBOL community could never have realized without financial support from the University of Oulu (2011), Kone foundation (2011–2013), Finnish Cultural Foundation (2012–2014), Finnish Ministry of the Environment (2012), and from 2014, the Academy of Finland through the Finnish Biodiversity Information Facility (FinBIF) infrastructure project on national roadmap. Colleagues in the iBOL community, particularly representatives of the national barcoding initiatives, have provided irreplaceable insights and experience in helping us to steer FinBOL in the right direction. Finally, we thank the Canada Foundation for Innovation, Canada First Research Excellence Fund, Genome Canada, Ontario Genomics, Ontario Ministry of Research and Innovation, and the Ontario Research Excellence Fund for supporting the iBOL consortium, the Centre for Biodiversity Genomics and the BOLD database which played an essential role in enabling both international and national barcoding campaigns. TR and OO were funded by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 856506; ERC-synergy project LIFEPLAN). OO was funded by Academy of Finland (grant no. 309581), Jane and Aatos Erkko Foundation, Research Council of Norway through its Centres of Excellence Funding Scheme (223257). Manuel Frías provided extensive, invaluable help in creating the figures in this manuscript.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

AUTHOR CONTRIBUTIONS

Marko Mutanen has served as the leader of the FinBOL initiative since its launch. Panu Somervuo implemented FinPROTAX, validated its performance and analysed the environmental sample. MP extracted and curated the final data sets. Tomas Roslin, Panu Somervuo, Marko Mutanen and Mikko Pentinsaari wrote the first draft of the manuscript, and all authors contributed to later revisions. The authors declare no conflict of interest.

DATA AVAILABILITY STATEMENT

All sequences, trace files and specimen metadata are publicly available in BOLD. The sequences of the FinBOL arthropod reference library are available as dx.doi.org/10.5883/DS-FINPRO, whereas the environmental sample of Finnish insects (GMTPC) is available as dx.doi.org/10.5883/DS-GMTPC. FinPROTAX software can be accessed at https://github.com/psomervuo/FinPROTAX and a web interface for FinPROTAX is available at https://laji.fi/en/theme/protax.

ORCID

Tomas Roslin https://orcid.org/0000-0002-2957-4791
Jeremy deWaard https://orcid.org/0000-0001-9778-5454
Vladislav Ivanov https://orcid.org/0000-0001-5783-950X
Janne Koskinen https://orcid.org/0000-0002-5396-575X
Mikko Tiusanen https://orcid.org/0000-0002-9361-0777
Eero J. Vesterinen https://orcid.org/0000-0003-3665-5802
Marko Mutanen https://orcid.org/0000-0003-4464-6308

REFERENCES

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, *215*(3), 403–410. https://doi.org/10.1016/S0022-2836(05)80360-2

Andersen, J. C., Oboyski, P., Davies, N., Charlat, S., Ewing, C., Meyer, C., Krehenwinkel, H., Lim, J. Y., Noriyuki, S., Ramage, T., Gillespie, R. G., & Roderick, G. K. (2019). Categorization of species as native or nonnative using DNA sequence signatures without a complete reference library. *Ecological Applications*, *29*(5), e01914. https://doi.org/10.1002/eap.1914

Antão, L. H., Pöyry, J., Leinonen, R., & Roslin, T. (2020). Contrasting latitudinal patterns in diversity and stability in a high-latitude

species-rich moth community. *Global Ecology and Biogeography*, *29*(5), 896–907. https://doi.org/10.1111/geb.13073

Barsoum, N., Bruce, C., Forster, J., Ji, Y. Q., & Yu, D. W. (2019). The devil is in the detail: Metabarcoding of arthropods provides a sensitive measure of biodiversity response to forest stand composition compared with surrogate measures of biodiversity. *Ecological Indicators*, *101*, 313–323. https://doi.org/10.1016/j.ecolind.2019.01.023

Bergsten, J., Bilton, D. T., Fujisawa, T., Elliott, M., Monaghan, M. T., Balke, M., Hendrich, L., Geijer, J., Herrmann, J., Foster, G. N., Ribera, I., Nilsson, A. N., Barraclough, T. G., & Vogler, A. P. (2012). The effect of geographical scale of sampling on DNA barcoding. *Systematic Biology*, *61*(5), 851–869. https://doi.org/10.1093/sysbio/sys037

Blaxter, M., Mann, J., Chapman, T., Thomas, F., Whitton, C., Floyd, R., & Abebe, E. (2005). Defining operational taxonomic units using DNA barcode data. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *360*(1462), 1935–1943. https://doi.org/10.1098/rstb.2005.1725

BOLD Team (2019). *Barcode of Life Data Systems Handbook – A web-based bioinformatics platform supporting the DNA barcoding of animal, plant, and fungal species*. Retrieved from www.boldsystems.org. version 4.0 (Draft 1).

Boonstra, H., Rinne, A., Kubiak, M., & Wiberg-Larsen, P. (2018). Description of the larva of *Holocentropus insignis martynov* 1924 (Trichoptera: Polycentropodidae) with notes on biology and distribution. *Zootaxa*, *4532*(2), 231–247. https://doi.org/10.11646/zootaxa.4532.2.3

Callahan, B. J., McMurdie, P. J., Rosen, M. J., Han, A. W., Johnson, A. J. A., & Holmes, S. P. (2016). DADA2: High-resolution sample inference from Illumina amplicon data. *Nature Methods*, *13*(7), 581–583. https://doi.org/10.1038/nmeth.3869

Clare, E. L., Fazekas, A. J., Ivanova, N. V., Floyd, R. M., Hebert, P. D. N., Adams, A. M., Nagel, J., Girton, R., Newmaster, S. G., & Fenton, M. B. (2019). Approaches to integrating genetic data into ecological networks. *Molecular Ecology*, *28*(2), 503–519. https://doi.org/10.1111/mec.14941

deWaard, J. R., Levesque-Beaudin, V., DeWaard, S. L., Ivanova, N. V., McKeown, J. T. A., Miskie, R., Naik, S., Perez, K. H. J., Ratnasingham, S., Sobel, C. N., Sones, J. E., Steinke, C., Telfer, A. C., Young, A. D., Young, M. R., Zakharov, E. V., & Hebert, P. D. N. (2019). Expedited assessment of terrestrial arthropod diversity by coupling Malaise traps with DNA barcoding. *Genome*, *62*(3), 85–95. https://doi.org/10.1139/gen-2018-0093

Dincă, V., Dapporto, L., Somervuo, P., Vodă, R., Cuvelier, S., Gascoigne-Pees, M., Huemer, P., Mutanen, M., Hebert, P. D. N., & Vila, R. (2021). High resolution DNA barcode library for European butterflies reveals continental patterns of mitochondrial genetic diversity. *Communications Biology*, *4*(1), 1–11.

Donner, J. (1995). *The quaternary history of scandinavia*. Cambridge University Press.

Ebach, M. C., & Holdrege, C. (2005). 2005 DNA barcoding is no substitute for taxonomy. *Nature*, *434*, 697. https://doi.org/10.1038/434697b

FinBIF (2020). *The FinBIF checklist of Finnish species 2019*. Finnish Biodiversity Information Facility, Finnish Museum of Natural History, University of Helsinki. Retrieved from http://urn.fi/URN:ISSN:2490-0907.

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*(5), 294–299. https://doi.org/10.1371/journal.pone.0013102

Geiger, M. F., Moriniere, J., Hausmann, A., Haszprunar, G., Wägele, W., Hebert, P. D. N., & Rulik, B. (2016). Testing the Global Malaise Trap Program – How well does the current barcode reference library identify flying insects in Germany? *Biodiversity Data Journal*, *4*(1), e10671. https://doi.org/10.3897/BDJ.4.e10671

Geller, J., Meyer, C., Parker, M., & Hawk, H. (2013). Redesign of PCR primers for mitochondrial cytochrome c oxidase subunit I for marine invertebrates and application in all-taxa biotic surveys. *Molecular Ecology Resources*, *13*(5), 851–861. https://doi.org/10.1111/1755-0998.12138

Haarto, A., & Ståhls, G. (2014). When mtDNA COI is misleading: Congruent signal of ITS2 molecular marker and morphology for north European Melanostoma Schiner, 1860 (Diptera, Syrphidae). *ZooKeys*, *431*, 93–134. https://doi.org/10.3897/zookeys.431.7207

Hausmann, A., Charles, H., Godfray, J., Huemer, P., Mutanen, M., Rougerie, R., Van Nieukerken, E. J., Ratnasingham, S., & Hebert, P. D. N. (2013). Genetic patterns in European geometrid moths revealed by the Barcode Index Number (BIN) system. *PLoS One*, *8*(12), e84518. https://doi.org/10.1371/journal.pone.0084518

Hausmann, A., Miller, S. E., Holloway, J. D., Dewaard, J. R., Pollock, D., Prosser, S. W. J., & Hebert, P. D. N. (2016). Calibrating the taxonomy of a megadiverse insect family: 3000 DNA barcodes from geometrid type specimens (Lepidoptera, Geometridae). *Genome*, *59*(9), 671–684. https://doi.org/10.1139/gen-2015-0197

Hebert, P. D. N., Braukmann, T. W. A., Prosser, S. W. J., Ratnasingham, S., deWaard, J. R., Ivanova, N. V., Janzen, D. H., Hallwachs, W., Naik, S., Sones, J. E., & Zakharov, E. V. (2018). A Sequel to Sanger: Amplicon sequencing that scales. *BMC Genomics 19*, 1–14. https://doi.org/10.1186/s12864-018-4611-3

Hebert, P. D. N., Cywinska, A., Ball, S. L., & DeWaard, J. R. (2003). Biological identifications through DNA barcodes. *Proceedings of the Royal Society B: Biological Sciences*, *270*(1512), 313–321. https://doi.org/10.1098/rspb.2002.2218

Hebert, P. D. N., & Gregory, T. R. (2005). The promise of DNA barcoding for taxonomy. *Systematic Biology*, *54*(5), 852–859. https://doi.org/10.1080/10635150500354886

Hebert, P. D. N., Hollingsworth, P. M., & Hajibabaei, M. (2016). From writing to reading the encyclopedia of life. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *371*(1702), 20150321. https://doi.org/10.1098/rstb.2015.0321

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*. *Proceedings of the National Academy of Sciences of the United States of America*, *101*(41), 14812–14817. https://doi.org/10.1073/pnas.0406166101

Hebert, P. D. N., Ratnasingham, S., Zakharov, E. V., Telfer, A. C., Levesque-Beaudin, V., Milton, M. A., Pedersen, S., Jannetta, P., & Dewaard, J. R. (2016). Counting animal species with DNA barcodes: Canadian insects. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *371*(1702), 20150333. https://doi.org/10.1098/rstb.2015.0333

Heikkilä, M., Mutanen, M., Kekkonen, M., & Kaila, L. (2014). Morphology reinforces proposed molecular phylogenetic affinities: A revised classification for Gelechioidea (Lepidoptera). *Cladistics*, *30*(6), 563–589. https://doi.org/10.1111/cla.12064

Hillebrand, H. (2004). On the generality of the latitudinal diversity gradient. *American Naturalist*, *163*(2), 192–211. https://doi.org/10.1086/381004

Hovmöller, R., Forshage, M., & Ronquist, F. (2017). *Streckkodning av svenska floran och faunan – förutsättningar och utmaningar*. PM från Naturhistoriska riksmuseet. 2017:1. Retrieved from http://www.nrm.se/download/18.2a85cf3215e50a71056eac46/1511363009228/Streckkodningsrapport_slutversion_2017-11-21.pdf.

Huemer, P., Hebert, P. D. N., Mutanen, M., Wieser, C., Wiesmair, B., Hausmann, A., Yakovlev, R., Möst, M., Gottsberger, B., Strutzenberger, P., & Fiedler, K. (2018). Large geographic distance versus small DNA barcode divergence: Insights from a comparison of European to South Siberian Lepidoptera. *PLoS One*, *13*(11), e0206668. https://doi.org/10.1371/journal.pone.0206668

Huemer, P., Karsholt, O., Aarvik, L., Berggren, K., Bidzilya, O., Junnilainen, J., Landry, J. F., Mutanen, M., Nupponen, K., Segerer, A., Šumpich,

J., Wieser, C., Wiesmair, B., & Hebert, P. D. N. (2020). DNA barcode library for European Gelechiidae (Lepidoptera) suggests greatly underestimated species diversity. *ZooKeys*, *2020*(921), 141–157. https://doi.org/10.3897/zookeys.921.49199

Huemer, P., & Mutanen, M. (2015). Alpha taxonomy of the genus *Kessleria* Nowicki, 1864, revisited in light of DNA-barcoding (Lepidoptera, Yponomeutidae). *ZooKeys*, *503*, 89–133. https://doi.org/10.3897/zookeys.503.9590

Huemer, P., Mutanen, M., Sefc, K. M., & Hebert, P. D. N. (2014). Testing DNA barcode performance in 1000 species of European Lepidoptera: Large geographic distances have small genetic impacts. *PLoS One*, *9*(12), https://doi.org/10.1371/journal.pone.0115774

Ivanov, V., Lee, K. M., & Mutanen, M. (2018). Mitonuclear discordance in wolf spiders: genomic evidence for species integrity and introgression. *Molecular Ecology*, *27*, 1681–1695. https://doi.org/10.1111/mec.14564

Ivanova, N.V., Dewaard, J.R., & Hebert, P.D.N. (2006). An inexpensive, automation-friendly protocol for recovering high-quality DNA. *Molecular Ecology Notes*, *6*(4), 998–1002.

Janzen, D., & Hallwachs, W. (2019). How a tropical country can DNA barcode itself. *IBOL Barcode Bulletin*, *9*(1). https://doi.org/10.21083/ibol.v9i1.5526

Ji, Y., Huotari, T., Roslin, T., Schmidt, N. M., Wang, J., Yu, D. W., & Ovaskainen, O. (2020). SPIKEPIPE: A metagenomic pipeline for the accurate quantification of eukaryotic species occurrences and intraspecific abundance change using DNA barcodes or mitogenomes. *Molecular Ecology Resources*, *20*(1), 256–267. https://doi.org/10.1111/1755-0998.13057

Kaartinen, R., Stone, G. N., Hearn, J., Lohse, K., & Roslin, T. (2010). Revealing secret liaisons: DNA barcoding changes our understanding of food webs. *Ecological Entomology*, *35*(5), 623–638. https://doi.org/10.1111/j.1365-2311.2010.01224.x

Kaila, L., Nupponen, K., Gorbunov, P. Y., Mutanen, M., & Heikkilä, M. (2020). Ustyurtiidae, a new family of Urodoidea with description of a new genus and two species from Kazakhstan, and discussion on possible affinity of Urodoidea to Schreckensteinioidea (Lepidoptera). *Insect Systematics and Evolution*, *51*(3), 444–471. https://doi.org/10.1163/1876312X-00002209

Karsholt, O., Mutanen, M., Lee, S., & Kaila, L. (2013). A molecular analysis of the Gelechiidae (Lepidoptera, Gelechioidea) with an interpretative grouping of its taxa. *Systematic Entomology*, *38*, 334–348. https://doi.org/10.1111/syen.12006

Kekkonen, M., Mutanen, M., Kaila, L., Nieminen, M., & Hebert, P. D. N. (2015). Delineating species with DNA barcodes: A case of taxon dependent method performance in moths. *PLoS One*, *10*(4), e0122481. https://doi.org/10.1371/journal.pone.0122481

Keret, N. M., Mutanen, M. J., Orell, M. I., Itämies, J. H., & Välimäki, P. M. (2020). Climate change-driven elevational changes among boreal nocturnal moths. *Oecologia*, *192*(4), 1085–1098. https://doi.org/10.1007/s00442-020-04632-w

Kirichenko, N., Triberti, P., Mutanen, M., Magnoux, E., Landry, J. F., & Lopez-Vaamonde, C. (2016). Systematics and biology of some species of *Micrurapteryx* Spuler (Lepidoptera, Gracillariidae) from the Holarctic Region, with re-description of *M. caraganella* (Hering) from Siberia. *ZooKeys*, *2016*(579), 99–156. https://doi.org/10.3897/zookeys.579.7166

Kivelä, S. M., Davis, R. B., Esperk, T., Gotthard, K., Mutanen, M., Valdma, D., & Tammaru, T. (2020). Comparative analysis of larval growth in Lepidoptera reveals instar-level constraints. *Functional Ecology*, *34*(7), 1391–1403. https://doi.org/10.1111/1365-2435.13556

Korpelainen, H., & Pietiläinen, M. (2017). Diversity of indoor fungi as revealed by DNA metabarcoding. *Genome*, *60*, 55–64. https://doi.org/10.1139/gen-2015-0191

Korpelainen, H., & Pietiläinen, M. (2019). The effects of sample age and taxonomic origin on the success rate of DNA barcoding when using herbarium material. *Plant Systematics and Evolution*, *305*(4), 319–324. https://doi.org/10.1007/s00606-019-01568-4

Korpelainen, H., Pietiläinen, M., & Huotari, T. (2016). Effective detection of indoor fungi by metabarcoding. *Annals of Microbiology*, *66*(1), 495–498. https://doi.org/10.1007/s13213-015-1118-x

Kozlov, M. V., Mutanen, M., Lee, K. M., & Huemer, P. (2017). Cryptic diversity in the long-horn moth *Nemophora degeerella* (Lepidoptera: Adelidae) revealed by morphology, DNA barcodes and genome-wide ddRAD-seq data. *Systematic Entomology*, *42*(2), 329–346. https://doi.org/10.1111/syen.12216

Landry, J. F., Nazari, V., Dewaard, J. R., Mutanen, M., Lopez-Vaamonde, C., Huemer, P., & Hebert, P. D. N. (2013). Shared but overlooked: 30 species of holarctic microlepidoptera revealed by DNA barcodes and morphology. *Zootaxa*, *3749*(1), 1–93. https://doi.org/10.11646/zootaxa.3749.1.1

Landvik, M., Wahlberg, N., Roslin, T., & Wahlberg, N. (2013). The identity of the Finnish *Osmoderma* (Coleoptera: Scarabaeidae, Cetoniinae) population established by COI sequencing. *Entomologica Fennica*, *24*, 147–155.

Lee, K. M., Kivelä, S. M., Ivanov, V., Hausmann, A., Kaila, L., Wahlberg, N., & Mutanen, M. (2018). Information dropout patterns in restriction site associated DNA phylogenomics and a comparison with multilocus Sanger data in a species-rich moth genus. *Systematic Biology*, *67*(6), 925–939. https://doi.org/10.1093/sysbio/syy029

Lee, K. M., Zeegers, T., Mutanen, M., & Pohjoismäki, J. (2020). The thin red line between species – genomic differentiation of *Gymnosoma* Meigen, a taxonomically challenging genus of parasitoid flies (Diptera: Tachinidae). *Systematic Entomology*, *46*, 96–110. https://doi.org/10.1111/syen.12450

Leray, M., Yang, J. Y., Meyer, C. P., Mills, S. C., Agudelo, N., Ranwez, V., Boehm, J. T., & Machida, R. J. (2013). A new versatile primer set targeting a short fragment of the mitochondrial COI region for metabarcoding metazoan diversity: Application for characterizing coral reef fish gut contents. *Frontiers in Zoology*, *10*(1), 34. https://doi.org/10.1186/1742-9994-10-34

Liston, A., Mutanen, M., & Viitasaari, M. (2019). On the taxonomy of *Heterarthrus* (Hymenoptera, Tenthredinidae), with a review of the West Palaearctic species. *Journal of Hymenoptera Research*, *72*, 83–126. https://doi.org/10.3897/jhr.72.39339

Lopez-Vaamonde, C., Sire, L., Rasmussen, B., Rougerie, R., Wieser, C., Allaoui, A. A., Minet, J., DeWaard, J. R., Decaëns, T., & Lees, D. C. (2019). DNA barcodes reveal deeply neglected diversity and numerous invasions of micromoths in Madagascar. *Genome*, *62*(3), 108–121. https://doi.org/10.1139/gen-2018-0065

Malaise, R. (1937). A new insect-trap. *Entomologisk Tidskrift*, *58*, 148–160.

Meierotto, S., Sharkey, M. J., Janzen, D. H., Hallwachs, W., Hebert, P. D. N., Chapman, E. G., & Smith, M. A. (2019). A revolutionary protocol to describe understudied hyperdiverse taxa and overcome the taxonomic impediment. *Deutsche Entomologische Zeitschrift*, *66*(2), 119–145. https://doi.org/10.3897/dez.66.34683

Meiklejohn, K. A., Damaso, N., & Robertson, J. M. (2019). Assessment of BOLD and GenBank – Their accuracy and reliability for the identification of biological materials. *PLoS One*, *14*(6), https://doi.org/10.1371/journal.pone.0217084

Miller, S. E., Hausmann, A., Hallwachs, W., & Janzen, D. H. (2016). Advancing taxonomy and bioinventories with DNA barcodes. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *371*(1702), 20150339. https://doi.org/10.1098/rstb.2015.0339

Miraldo, A., Krell, F. T., Smalén, M., Angus, R. B., & Roslin, T. (2014). Making the cryptic visible resolving the species complex of *Aphodius fimetarius* (Linnaeus) and *Aphodius pedellus* (de Geer) (Coleoptera: Scarabaeidae) by three complementary methods. *Systematic Entomology*, *39*(3), 531–547. https://doi.org/10.1111/syen.12079

Morinière, J., Balke, M., Doczkal, D., Geiger, M. F., Hardulak, L. A., Haszprunar, G., Hausmann, A., Hendrich, L., Regalado, L., Rulik,

B., Schmidt, S., Wägele, J. W., & Hebert, P. D. N. (2019). A DNA barcode library for 5,200 German flies and midges (Insecta: Diptera) and its implications for metabarcoding-based biomonitoring. *Molecular Ecology Resources*, *19*(4), 900–928. https://doi.org/10.1111/1755-0998.13022

Mutanen, M., Aarvik, L., Huemer, P., Kaila, L., Karsholt, O., & Tuck, K. (2012). DNA barcodes reveal that the widespread European tortricid moth *Phalonidia manniana* (Lepidoptera: Tortricidae) is a mixture of two species. *Zootaxa*, *3262*, 1–21. https://doi.org/10.11646/zootaxa.3262.1.1

Mutanen, M., Aarvik, L., Landry, J.-F., Segerer, A. H., & Karsholt, O. (2012). *Epinotia cinereana* (Haworth, 1811) bona sp., a Holarctic tortricid distinct from *E. nisella* (Clerck, 1759) (Lepidoptera: Tortricidae: Eucosmini) as evidenced by DNA barcodes, morphology and life history. *Zootaxa*, *3318*(1), 1–25.

Mutanen, M., Hausmann, A., Hebert, P. D. N., Landry, J. F., de Waard, J. R., & Huemer, P. (2012). Allopatry as a Gordian knot for taxonomists: Patterns of DNA barcode divergence in Arctic-Alpine Lepidoptera. *PLoS One*, *7*(10), e47214. https://doi.org/10.1371/journal.pone.0047214

Mutanen, M., Huemer, P., Autto, J., Karsholt, O., & Kaila, L. (2020). *Monopis jussii*, a new species (Lepidoptera, Tineidae) inhabiting nests of the boreal owl (*Aegolius funereus*). *ZooKeys*, *2020*(992), 157–181. https://doi.org/10.3897/zookeys.992.53975

Mutanen, M., Kaila, L., & Tabell, J. (2013). Wide-ranging barcoding aids discovery of one-third increase of species richness in presumably well-investigated moths. *Scientific Reports*, *3*, 1–7. https://doi.org/10.1038/srep02901

Mutanen, M., Kekkonen, M., Prosser, S. W. J., Hebert, P. D. N., & Kaila, L. (2015). One species in eight: DNA barcodes from type specimens resolve a taxonomic quagmire. *Molecular Ecology Resources*, *15*(4), 967–984. https://doi.org/10.1111/1755-0998.12361

Mutanen, M., Kivelä, S. M., Vos, R. A., Doorenweerd, C., Ratnasingham, S., Hausmann, A., Huemer, P., Dincă, V., van Nieukerken, E. J., Lopez-Vaamonde, C., Vila, R., Aarvik, L., Decaëns, T., Efetov, K. A., Hebert, P. D. N., Johnsen, A., Karsholt, O., Pentinsaari, M., Rougerie, R., … Godfray, H. C. J. (2016). Species-level para- and polyphyly in DNA barcode gene trees: Strong operational bias in European Lepidoptera. *Systematic Biology*, *65*(6), 1024–1040. https://doi.org/10.1093/sysbio/syw044

Mutanen, M., Ovaskainen, O., Várkonyi, G., Itämies, J., Prosser, S. W. J., Hebert, P. D. N., & Hanski, I. (2020). Dynamics of a host–parasitoid interaction clarified by modelling and DNA sequencing. *Ecology Letters*, *23*, 851–859. https://doi.org/10.1111/ele.13486

Nieminen, M., Hansson, C., Kekkonen, M., & Vikberg, V. (2018). *Mesopolobus incultus* auct. (Hymenoptera: Pteromalidae) contains two distinct species: *Mesopolobus incultus* (Walker, 1834) and *Mesopolobus amyntor* (Walker, 1845). *Entomologica Fennica*, *29*, 175–184. https://doi.org/10.33338/ef.77303

Nyman, T., Leppänen, S. A., Várkonyi, G., Shaw, M. R., Koivisto, R., Barstad, T. E., Vikberg, V., & Roininen, H. (2015). Determinants of parasitoid communities of willow-galling sawflies: Habitat overrides physiology, host plant and space. *Molecular Ecology*, *24*(19), 5059–5074. https://doi.org/10.1111/mec.13369

Ondov, B. D., Bergman, N. H., & Phillippy, A. M. (2011). Interactive metagenomic visualization in a Web browser. *BMC Bioinformatics*, *12*(1), 1–10. https://doi.org/10.1186/1471-2105-12-385

Padial, J. M., & De La Riva, I. (2007). Integrative taxonomists should use and produce DNA barcodes. *Zootaxa*, *68*(1586), 67–68. https://doi.org/10.11646/zootaxa.1586.1.7

Paukkunen, J., & Kozlov, M. V. (2015). Stinging wasps, ants and bees (Hymenoptera: Aculeata) of the Murmansk region, Northwest Russia. *Entomologica Fennica*, *26*, 53–73. https://doi.org/10.33338/ef.51282

Pentinsaari, M., Hebert, P. D. N., & Mutanen, M. (2014). Barcoding beetles: A regional survey of 1872 species reveals high identification success and unusually deep interspecific divergences. *PLoS One*, *9*(9), e108651. https://doi.org/10.1371/journal.pone.0108651

Pentinsaari, M., Mutanen, M., & Kaila, L. (2014). Cryptic diversity and signs of mitochondrial introgression in the *Agrilus viridis* species complex (Coleoptera: Buprestidae). *European Journal of Entomology*, *111*(4), 475–486. https://doi.org/10.14411/eje.2014.072

Pentinsaari, M., Ratnasingham, S., Miller, S. E., & Hebert, P. D. N. (2020). BOLD and GenBank revisited – Do identification errors arise in the lab or in the sequence libraries? *PLoS One*, *15*(4), 1–10. https://doi.org/10.1371/journal.pone.0231814

Pentinsaari, M., Salmela, H., Mutanen, M., & Roslin, T. (2016). Molecular evolution of a widely-adopted taxonomic marker (COI) across the animal tree of life. *Scientific Reports*, *6*(1), 1–12. https://doi.org/10.1038/srep35275

Pentinsaari, M., Vos, R., & Mutanen, M. (2017). Algorithmic single-locus species delimitation: Effects of sampling effort, variation and nonmonophyly in four methods and 1870 species of beetles. *Molecular Ecology Resources*, *17*(3), 393–404. https://doi.org/10.1111/1755-0998.12557

Pilipenko, V. E., Salmela, J., & Vesterinen, E. J. (2012). Description and DNA barcoding of *Tipula* (*Pterelachisus*) *recondita* sp. n. from the Palaearctic region (Diptera, Tipulidae). *ZooKeys*, *192*, 51–65. https://doi.org/10.3897/zookeys.192.2364

Pohjoismäki, J., & Haarto, A. (2015). *Linnaemya bergstroemi* n. sp. (Diptera: Tachinidae) – A new parasitoid fly from the Finnish Lapland. *Zootaxa*, *4059*(3), 581–597. https://doi.org/10.11646/zootaxa.4059.3.9

Pohjoismäki, J. L. O., Kahanpää, J., & Mutanen, M. (2016). DNA barcodes for the northern European tachinid flies (Diptera: Tachinidae). *PLoS One*, *11*(11), e0164933. https://doi.org/10.1371/journal.pone.0164933

Prosser, S. W. J., Dewaard, J. R., Miller, S. E., & Hebert, P. D. N. (2016). DNA barcodes from century-old type specimens using next-generation sequencing. *Molecular Ecology Resources*, *16*(2), 487–497. https://doi.org/10.1111/1755-0998.12474

Prous, M., Lee, K. M., & Mutanen, M. (2020). Cross-contamination and strong mitonuclear discordance in *Empria* sawflies (Hymenoptera, Tenthredinidae) in the light of phylogenomic data. *Molecular Phylogenetics and Evolution*, *143*, 106670. https://doi.org/10.1016/j.ympev.2019.106670

Prous, M., Vikberg, V., Liston, A., & Kramp, K. (2016). North-western Palaearctic species of the *Pristiphora ruficornis* group (Hymenoptera, Tenthredinidae). *Journal of Hymenoptera Research*, *51*, 1–54. https://doi.org/10.3897/jhr.51.9162

Pykälä, J., & Myllys, L. (2016). Three new species of *Atla* from calcareous rocks (Verrucariaceae, lichenized Ascomycota). *Lichenologist*, *48*(2), 111–120. https://doi.org/10.1017/S0024282915000523

Ratnasingham, S., & Hebert, P. D. N. (2007). BOLD: The barcode of life data system (www.barcodinglife.org). *Molecular Ecology Notes*, *7*, 355–364.

Ratnasingham, S., & Hebert, P. D. N. (2013). A DNA-based registry for all animal species: The Barcode Index Number (BIN) system. *PLoS One*, *8*(7), e66213. https://doi.org/10.1371/journal.pone.0066213

Ruggiero, A., & Werenkraut, V. (2007). One-dimensional analyses of Rapoport's rule reviewed through meta-analysis. *Global Ecology and Biogeography*, *16*(4), 401–414. https://doi.org/10.1111/j.1466-8238.2006.00303.x

Rytkönen, S., Vesterinen, E. J., Westerduin, C., Leviäkangas, T., Vatka, E., Mutanen, M., Välimäki, P., Hukkanen, M., Suokas, M., & Orell, M. (2019). From feces to data: A metabarcoding method for analyzing consumed and available prey in a bird-insect food web. *Ecology and Evolution*, *9*(1), 631–639. https://doi.org/10.1002/ece3.4787

Salmela, J. (2012). Biogeographic patterns of Finnish crane flies (Diptera, Tipuloidea). *Psyche (London)*, *2012*, 1–20. https://doi.org/10.1155/2012/913710

Salmela, J., Kaunisto, K. M., & Vahtera, V. (2014). Unveiling of a cryptic *Dicranomyia* (*Idiopyga*) from northern Finland using integrative

approach (Diptera, Limoniidae). *Biodiversity Data Journal*, *2*(1), https://doi.org/10.3897/BDJ.2.e4238

Schemske, D. W., & Mittelbach, G. G. (2017). "Latitudinal gradients in species diversity": Reflections on Pianka's 1966 article and a look forward. *American Naturalist*, *189*(6), 599–603. https://doi.org/10.1086/691719

Sihvonen, P., Lee, K. M., Lundsten, K. E., & Mutanen, M. (2020). Genomic evidence suggests *Mesapamea remmi* is an imaginary species (Lepidoptera: Noctuidae). *Systematic Entomology*, *45*(2), 302–311. https://doi.org/10.1111/syen.12397

Smith, A. M., Fernández-Triana, J. L., Eveleigh, E., Gómez, J., Guclu, C., Hallwachs, W., Hebert, P. D. N., Hrcek, J., Huber, J. T., Janzen, D., Mason, P. G., Miller, S., Quicke, D. L. J., Rodriguez, J. J., Rougerie, R., Shaw, M. R., Várkonyi, G., Ward, D. F., Whitfield, J. B., & Zaldívar-Riverón, A. (2013). DNA barcoding and the taxonomy of Microgastrinae wasps (Hymenoptera, Braconidae): Impacts after 8 years and nearly 20 000 sequences. *Molecular Ecology Resources*, *13*(2), 168–176. https://doi.org/10.1111/1755-0998.12038

Somervuo, P., Koskela, S., Pennanen, J., Nilsson, R. H., & Ovaskainen, O. (2016). Unbiased probabilistic taxonomic classification for DNA barcoding. *Bioinformatics*, *32*(19), 2920–2927. https://doi.org/10.1093/bioinformatics/btw346

Somervuo, P., Yu, D. W., Xu, C. C. Y., Ji, Y., Hultman, J., Wirta, H., & Ovaskainen, O. (2017). Quantifying uncertainty of taxonomic placement in DNA barcoding and metabarcoding. *Methods in Ecology and Evolution*, *8*(4), 398–407. https://doi.org/10.1111/2041-210X.12721

Ståhls, G., Miettinen, O., & Rättel, E. (2015). mtDNA COI in efficient use: Clarifying taxonomy, linking morphologically discordant sexes and identifying the immature stages of *Agathomyia* Verrall flat-footed flies (Diptera: Platypezidae). *Journal of Zoological Systematics and Evolutionary Research*, *53*(3), 219–238. https://doi.org/10.1111/jzs.12091

Stevens, G. (1996). Extending Rapoport's rule to Pacific marine fishes. *Journal of Biogeography*, *23*(2), 149–154. https://doi.org/10.1046/j.1365-2699.1996.00977.x

Tabell, J., Mutanen, M., & Siloaho, R. (2019). *Coleophora sirella* Tabeli & Mutanen, sp. N. from Finland (Lepidoptera: Coleophoridae). *Entomologica Fennica*, *30*(2), 49–56. https://doi.org/10.33338/ef.82918

Townes, H. (1972). A light-weight Malaise trap. *Entomology News*, *83*, 253–262.

van Nieukerken, E. J., Doorenweerd, C., Hoare, R. J. B., & Davis, D. R. (2016). Revised classification and catalogue of global Nepticulidae and Opostegidae (Lepidoptera, Nepticuloidea). *ZooKeys*, *2016*(628), 65–246. https://doi.org/10.3897/zookeys.628.9799

Van Nieukerken, E. J., Mutanen, M., & Doorenweerd, C. (2012). DNA barcoding resolves species complexes in *Stigmella salicis* and *S. aurella* species groups and shows additional cryptic speciation in *S. salicis* (Lepidoptera: Nepticulidae). *Entomologisk Tidskrift*, *132*(4), 235–255.

Vesterinen, E. J., Lilley, T., Laine, V. N., & Wahlberg, N. (2013). Next generation sequencing of fecal DNA reveals the dietary diversity of the widespread insectivorous predator Daubenton's Bat (*Myotis daubentonii*) in Southwestern Finland. *PLoS One*, *8*(11), e82168. https://doi.org/10.1371/journal.pone.0082168

Vesterinen, E. J., Ruokolainen, L., Wahlberg, N., Peña, C., Roslin, T., Laine, V. N., Vasko, V., Sääksjärvi, I. E., Norrdahl, K., & Lilley, T. M. (2016). What you need is what you eat? Prey selection by the bat *Myotis daubentonii*. *Molecular Ecology*, *25*(7), 1581–1594. https://doi.org/10.1111/mec.13564

Wikström, B., Huemer, P., Mutanen, M., Tyllinen, J., & Kaila, L. (2020). *Pyralis cardinalis*, a charismatic new species related to *P. regalis* [Denis & Schiffermüller], 1775, first recognized in Finland (Lepidoptera, Pyralidae). *Nota Lepidopterologica*, *43*, 337–364. https://doi.org/10.3897/nl.43.54916

Will, K. W., Mishler, B. D., & Wheeler, Q. D. (2005). The perils of DNA barcoding and the need for integrative taxonomy. *Systematic Biology*, *54*(5), 844–851. https://doi.org/10.1080/10635150500354878

Wilson, J. J., Rougerie, R., Schonfeld, J., Janzen, D. H., Hallwachs, W., Hajibabaei, M., Kitching, I. J., Haxaire, J., & Hebert, P. D. N. (2011). When species matches are unavailable are DNA barcodes correctly assigned to higher taxa? An assessment using sphingid moths. *BMC Ecology*, *11*, 18. https://doi.org/10.1186/1472-6785-11-18

Wirta, H., Várkonyi, G., Rasmussen, C., Kaartinen, R., Schmidt, N. M., Hebert, P. D. N., Barták, M., Blagoev, G., Disney, H., Ertl, S., Gjelstrup, P., Gwiazdowicz, D. J., Huldén, L., Ilmonen, J., Jakovlev, J., Jaschhof, M., Kahanpää, J., Kankaanpää, T., Krogh, P. H., … Roslin, T. (2016). Establishing a community-wide DNA barcode library as a new tool for arctic research. *Molecular Ecology Resources*, *16*(3), 809–822. https://doi.org/10.1111/1755-0998.12489

Wirta, H. K., Vesterinen, E. J., Hambäck, P. A., Weingartner, E., Rasmussen, C., Reneerkens, J., Schmidt, N. M., Gilg, O., & Roslin, T. (2015). Exposing the structure of an Arctic food web. *Ecology and Evolution*, *5*(17), 3842–3856. https://doi.org/10.1002/ece3.1647

Zhou, X., Frandsen, P. B., Holzenthal, R. W., Beet, C. R., Bennett, K. R., Blahnik, R. J., Bonada, N., Cartwright, D., Chuluunbat, S., Cocks, G. V., Collins, G. E., deWaard, J., Dean, J., Flint, O. S., Hausmann, A., Hendrich, L., Hess, M., Hogg, I. D., Kondratieff, B. C., … Kjer, K. M. (2016). The Trichoptera barcode initiative: A strategy for generating a species-level tree of life. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *371*(1702), 20160025. https://doi.org/10.1098/rstb.2016.0025

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

How to cite this article: Roslin, T., Somervuo, P., Pentinsaari, M., Hebert, P. D. N., Agda, J., Ahlroth, P., Anttonen, P., Aspi, J., Blagoev, G., Blanco, S., Chan, D., Clayhills, T., deWaard, J., deWaard, S., Elliot, T., Elo, R., Haapala, S., Helve, E., Ilmonen, J., … Mutanen, M. (2022). A molecular-based identification resource for the arthropods of Finland. *Molecular Ecology Resources*, 22, 803–822. https://doi.org/10.1111/1755-0998.13510