*Syst. Biol.* 71(6):1404–1422, 2022

https://doi.org/10.1093/sysbio/syac033
Advance Access publication May 13, 2022

# Towards Large-Scale Integrative Taxonomy (LIT): Resolving the Data Conundrum for Dark Taxa

Emily Hartop[1,2,3], Amrita Srivathsan[3,4], Fredrik Ronquist[5], and Rudolf Meier[3,4,*]

[1]*Department of Zoology, Stockholm University, Frescativägen, 114 19 Stockholm, Sweden;* [2]*Station Linné, Skogsby 161, 386 93 Färjestaden, Sweden;* [3]*Center for Integrative Biodiversity Discovery, Leibniz Institute for Evolution and Biodiversity Science, Museum für Naturkunde, Invalidenstraße 43, 10115 Berlin, Germany;* [4]*Department of Biological Sciences, National University of Singapore, 21 Lower Kent Ridge Rd., Singapore 119077, Singapore and* [5]*Department of Bioinformatics and Genetics, Swedish Museum of Natural History, Box 50007, SE-104 05 Stockholm, Sweden*

[*]*Correspondence to be sent to: Center for Integrative Biodiversity Discovery, Leibniz Institute for Evolution and Biodiversity Science, Museum für Naturkunde, Invalidenstraße 43, 10115 Berlin, Germany;*
*E-mail: Rudolf.Meier@mfn.berlin.*

*Received 10 May 2021; accepted 22 April 2022*
*Associate Editor: Benoit Dayrat*

*Abstract*.—New, rapid, accurate, scalable, and cost-effective species discovery and delimitation methods are needed for tackling "dark taxa," here defined as groups for which <10% of all species are described and the estimated diversity exceeds 1,000 species. Species delimitation for these taxa should be based on multiple data sources ("integrative taxonomy") but collecting multiple types of data risks impeding a discovery process that is already too slow. We here develop large-scale integrative taxonomy (LIT), an explicit method where preliminary species hypotheses are generated based on inexpensive data that can be obtained quickly and cost-effectively. These hypotheses are then evaluated based on a more expensive type of "validation data" that is only obtained for specimens selected based on objective criteria applied to the preliminary species hypotheses. We here use this approach to sort 18,000 scuttle flies (Diptera: Phoridae) into 315 preliminary species hypotheses based on next-generation sequencing barcode (313 bp) clusters (using objective clustering [OC] with a 3% threshold). These clusters are then evaluated with morphology as the validation data. We develop quantitative indicators for predicting which barcode clusters are likely to be incongruent with morphospecies by randomly selecting 100 clusters for in-depth validation with morphology. A linear model demonstrates that the best predictors for incongruence between barcode clusters and morphology are maximum p-distance within the cluster and a newly proposed index that measures cluster stability across different clustering thresholds. A test of these indicators using the 215 remaining clusters reveals that these predictors correctly identify all clusters that are incongruent with morphology. In our study, all morphospecies are true or disjoint subsets of the initial barcode clusters so that all incongruence can be eliminated by varying clustering thresholds. This leads to a discussion of when a third data source is needed to resolve incongruent grouping statements. The morphological validation step in our study involved 1,039 specimens (5.8% of the total). The formal LIT protocol we propose would only have required the study of 915 (5.1%: 2.5 specimens per species), as we show that clusters without signatures of incongruence can be validated by only studying two specimens representing the most divergent haplotypes. To test the generality of our results across different barcode clustering techniques, we establish that the levels of incongruence are similar across OC, Automatic Barcode Gap Discovery (ABGD), Poisson Tree Processes (PTP), and Refined Single Linkage (RESL) (used by Barcode of Life Data System to assign Barcode Index Numbers [BINs]). OC and ABGD achieved a maximum congruence score with the morphology of 89% while PTP was slightly less effective (84%). RESL could only be tested for a subset of the specimens because the algorithm is not public. BINs based on 277 of the original 1,714 haplotypes were 86% congruent with morphology while the values were 89% for OC, 74% for PTP, and 72% for ABGD. [Biodiversity discovery; dark taxa; DNA barcodes; integrative taxonomy.]

"I saw with regret, (and all scientific men have shared this feeling) that whilst the number of accurate instruments was daily increasing, we were still ignorant"

*Alexander von Humboldt*

In a recent report, global reinsurance giant Swiss Re concluded that "55% of global gross domestic product is moderately or highly dependent" on biodiversity and ecosystem services that "underpin all economic activity in our societies globally" (Schelske et al. 2020). Such services are critically dependent on functionally diverse invertebrate groups like insects that contribute a wide range of ecosystem services (Losey and Vaughan 2006), comprise over half of the described species (Chapman 2009), and are hosts for millions of unique bacterial species, nematodes, and mites (Larsen et al. 2017). Unfortunately, the vast majority of this diversity remains undescribed (Stork 2018), leaving humanity dangerously ignorant, with most species largely inaccessible for ecological studies and biodiversity inventories. This highlights the need for completing one of the great incomplete tasks in science—an inventory of all life (Mora et al. 2011; Stork 2018).

The bulk of the planet's unknown diversity is in hyperdiverse, taxonomically neglected groups. These taxa are often so diverse that a reasonably precise estimate of true species numbers is currently impossible, leading to their referral as "open-ended" (Bickel 2009) or, more recently, "dark" taxa. The latter was originally coined for the growing number of sequences in GenBank that were not linked to formal scientific names (Page 2011, 2016) but has evolved to refer to species-rich taxa of small body size for whom most of the species-level diversity is undescribed (e.g., Hausmann et al. 2020). Here, we not only accept the current usage but also propose that the term should only be applied to taxa for which the undescribed fauna is estimated to exceed the described fauna by at least one order of magnitude and the total diversity exceeds 1000 species. Descriptive work on these groups has been very slow in part because a single site can yield thousands of specimens belonging to hundreds of species (Puillandre et al. 2012; Srivathsan et al. 2019).

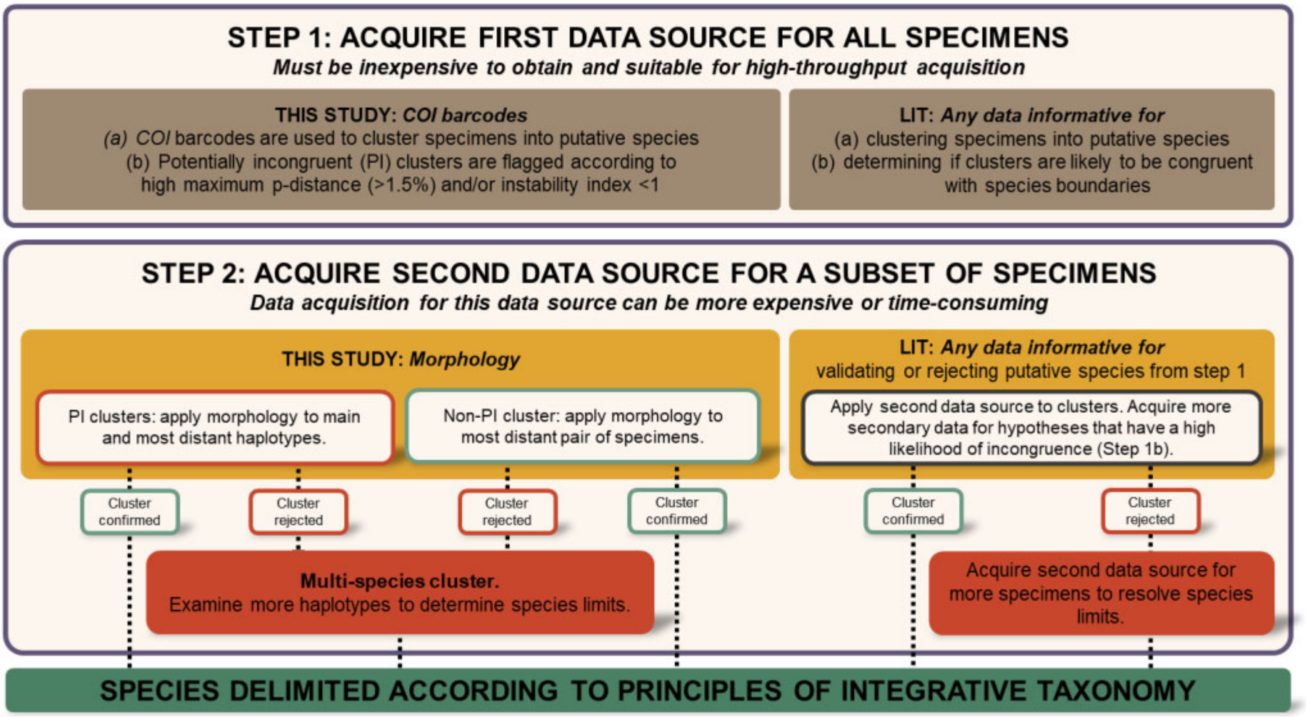
STEP 1: ACQUIRE FIRST DATA SOURCE FOR ALL SPECIMENS
Must be inexpensive to obtain and suitable for high-throughput acquisition
THIS STUDY: COI barcodes
(a) COI barcodes are used to cluster specimens into putative species
(b) Potentially incongruent (PI) clusters are flagged according to high maximum p-distance (>1.5%) and/or instability index <1
LIT: Any data informative for
(a) clustering specimens into putative species
(b) determining if clusters are likely to be congruent with species boundaries
STEP 2: ACQUIRE SECOND DATA SOURCE FOR A SUBSET OF SPECIMENS
Data acquisition for this data source can be more expensive or time-consuming
THIS STUDY: Morphology
PI clusters: apply morphology to main and most distant haplotypes.
Non-PI cluster: apply morphology to most distant pair of specimens.
LIT: Any data informative for
validating or rejecting putative species from step 1
Apply second data source to clusters. Acquire more secondary data for hypotheses that have a high likelihood of incongruence (Step 1b).
Cluster confirmed
Cluster rejected
Cluster rejected
Cluster confirmed
Cluster confirmed
Cluster rejected
Multi-species cluster.
Examine more haplotypes to determine species limits.
Acquire second data source for more specimens to resolve species limits.
SPECIES DELIMITED ACCORDING TO PRINCIPLES OF INTEGRATIVE TAXONOMY

FIGURE 1. LIT protocol. Two data sources are used: the first is collected for all specimens, the second for a select subset of specimens based on analysis of the primary data.

Tackling such samples with traditional morphological methods is painfully slow. Consequently, dark taxa are often either ignored entirely or only a few specimens are "cherry picked" for study.

The proposal of DNA barcoding for the identification of organisms (Hebert et al. 2003) and DNA taxonomy for the delimitation of organisms (Tautz et al. 2003; Blaxter 2004; Vogler and Monaghan 2007) in the early 2000s highlighted the potential of DNA sequences for accelerating species discovery, identification, delimitation, and description—especially in difficult groups. However, only recently have molecular methods become sufficiently cost-effective and robust for a true reversal of the traditional workflow that is based on first sorting specimens with morphology and then collecting DNA sequences for a select subset of specimens (Puillandre et al. 2012; Kekkonen and Hebert 2014; Wang et al. 2018; Yeo et al. 2020). Some authors have even proposed that the validation of barcode clusters with morphology or any other type of data is unnecessary, but a single short barcode contains only a limited amount of information relevant to species boundaries (Kwong et al. 2012; Pentinsaari et al. 2016, Meier et al. 2021). Indeed, recent analyses of large barcode data sets have revealed that 10–20% of all barcode clusters differ depending on the method and parameters used for molecular species delimitation (Meier et al. 2021) with some taxa displaying even higher levels of incongruence (Kekkonen and Hebert 2014; Meier et al. 2021). This means that multiple character systems are needed for accurately delimiting species (integrative taxonomy: Dayrat 2005; Padial et al. 2010; Schlick-Steiner et al. 2010; Puillandre et al. 2012; Ratnasingham and Hebert 2013; Zhang et al. 2013; Pante et al. 2015; Vitecek et al. 2017).

Despite this recognized need, efficient approaches to integrative species delimitation are still underdeveloped. One approach to large-scale integrative taxonomy using COI barcodes was proposed by Puillandre et al. (2012) wherein primary species hypotheses were derived using multiple molecular delimitation methods (Automated Barcode Gap Discovery and General Mixed Yule Coalescence Method), the cluster incongruence between methods was visualized, and other data sources (nuclear, morphological, and geographic) were used to test the primary hypotheses. This method yields high-quality species limits but its requirement of having to collect so much data for all specimens also slows down taxonomic progress that is already too slow for dark taxa. Therefore, the traditional approach to dark taxa has been morphospecies sorting followed by the barcoding of a few representatives without a tested set of rules for choosing which specimens should be barcoded. We herein propose LIT (Fig. 1) that includes a set of rules that are designed to minimize data collection while ensuring the use of at least two types of data per species. The core principle is first generating preliminary species-level hypotheses based on a data source that can be acquired quickly and in semiautomated ways. In our study, this goal is achieved by clustering COI barcodes, which can now be inexpensively and easily obtained due to the development of individual-specimen next-generation sequencing (NGS) techniques

(Hebert et al. 2018; Srivathsan et al. 2018, 2019, 2021; Wang et al. 2018). In future implementations of LIT, other data sources (e.g., high throughput, robotic imaging: Wührl et al. 2021) could replace or complement barcodes for this first step. For example, this will be necessary for some dark taxa where standard DNA barcodes may not be satisfying for species-level sorting (e.g., fungi, algae, etc.). We then show that there are systematic ways to evaluate the preliminary species hypotheses that are based on the first data source to determine those that are weak and require further scrutiny. This allows us to acquire the second data source for a limited number of specimens that are chosen to validate or reject the preliminary hypotheses.

In our study, we show that there are COI barcode cluster-specific traits that can predict whether a molecular cluster will be incongruent with species boundaries. We identify two such predictors through a thorough examination of 100 clusters and then confirm their effectiveness by testing them for the remaining 215 barcode clusters. This leads to the development of explicit rules for picking specimens for validating preliminary species hypotheses based on barcode clusters. These rules are formalized in an algorithm to demonstrate that the system is effective for our sample. LIT requires that weak hypotheses are more rigorously tested while strong species hypotheses are validated by collecting validation data for only two specimens. The second type of data can therefore be more expensive and/or require more highly skilled manpower because they only need to be acquired for a small number of specimens. In our study, the second data source is morphology, but one could equally well use nuclear markers or other data sources relevant for the species delimitation of the taxon in question.

We here demonstrate how LIT can be used to discover the species-level diversity of 18,000 specimens of scuttle flies (Diptera: Phoridae) from Sweden. Our application of LIT to Swedish phorids allowed us to solve a taxonomic problem that was too large for the application of traditional techniques, as historically these samples would have been ignored or only morphologically unusual species would have been "cherry picked" for study. We predict that future applications of LIT will follow the general principles outlined here but likely require customization and calibration of the criteria that are used to create the primary species hypotheses and to flag those hypotheses that have a high chance for being incongruent (Fig. 1).

## Materials and Methods

### Sampling

Samples were collected with Townes-style Malaise traps (Townes 1972) at 36 sites across Sweden as part of the Swedish Insect Inventory Project (Fig. 2a; Karlsson et al. 2020). A single sample from late spring/early summer 2018 (except for site 46, where the first available sample was from July) was selected from each site (Supplementary Table S1), and specimens from the dipteran family Phoridae were extracted for sequencing. Due to the high number of phorids present (numbering in the hundreds of thousands), only a randomly selected subsample of the specimens was sequenced. The specimens were kept in ethanol at −20 to 25°C until processing.

### *DNA Extraction, Polymerase Chain Reaction, and Sequencing*

DNA extractions were carried out nondestructively on whole flies using 10 μl of "HotSHOT" solution (Truett et al. 2000; Srivathsan et al. 2019). Incubation was in a thermocycler at 65°C for 15 min followed by 98°C for 2 min. A total of 206 96-well plates (19,570 specimens) were extracted. The DNA extracts were used to set up plates of polymerase chain reactions (PCRs) to amplify a 313-bp minibarcode fragment of the COI barcoding region using m1COlintF: 5′-GGW ACWGGWTGAACWGTWTAYCCYCC-3′ (Leray et al. 2013) and modified jgHCO2198: 50-TANACYTC NGGRTGNCCRAARAAYCA-3 (Geller et al. 2013). Amplifications were conducted with tagged primers and sequenced with Illumina HiSeq 2500 or Oxford Nanopore Technologies MinION. The Illumina data were analyzed following the protocols first established in Meier et al. (2016) and then modified for Wang et al. (2018). The processing of the MinION data followed the bioinformatics pipeline in Srivathsan et al. (2019), which was shown to yield barcodes that were essentially error-free based on comparing amplicon sequences obtained with Illumina and MinION for the same specimens (99.99% accuracy, 0.46% ambiguity). Note that the latest methods for MinION barcoding with a recent flowcell type (R10.3) yield the same accuracy but lower ambiguity levels (<0.01%: Srivathsan et al. 2021).

PCR reactions contained 4 μL Mastermix from CWBio, 1 μL of 1 mg/mL BSA, 1 μL of 10 μM of each primer, and 1 μL of DNA. PCR conditions were a 5 min initial denaturation at 94°C followed by 35 cycles of denaturation/annealing/extension (94°C [1 min]/47°C [2 min]/72°C [1 min]), and a final extension at 72°C (5 min). PCR products were pooled, cleaned, and sequenced in either a lane of HiSeq 2500 (250 bp paired-end sequencing) or a MinION R9.4 flowcell. For MinION, the SQK-LSK109 ligation sequencing kit (Oxford Nanopore Technologies) was used for preparing a library from 200-ng of the pooled and purified PCR products for sequencing. The manufacturer's instructions were followed except for the use of 1× instead of 0.4× Ampure beads (Beckmann Coulter) because the amplicons in our experiments were short (~391 bp with primers and tags). Illumina libraries were prepared using TruSeq DNA PCR-free kits to obtain 250 bp PE sequences using Illumina HiSeq 2500. Illumina sequencing was outsourced. Nanopore sequencing using a MinION sequencer was conducted in house following the description in Srivathsan et al. (2019), and basecalling was conducted using Guppy 2.3.5+53a111f

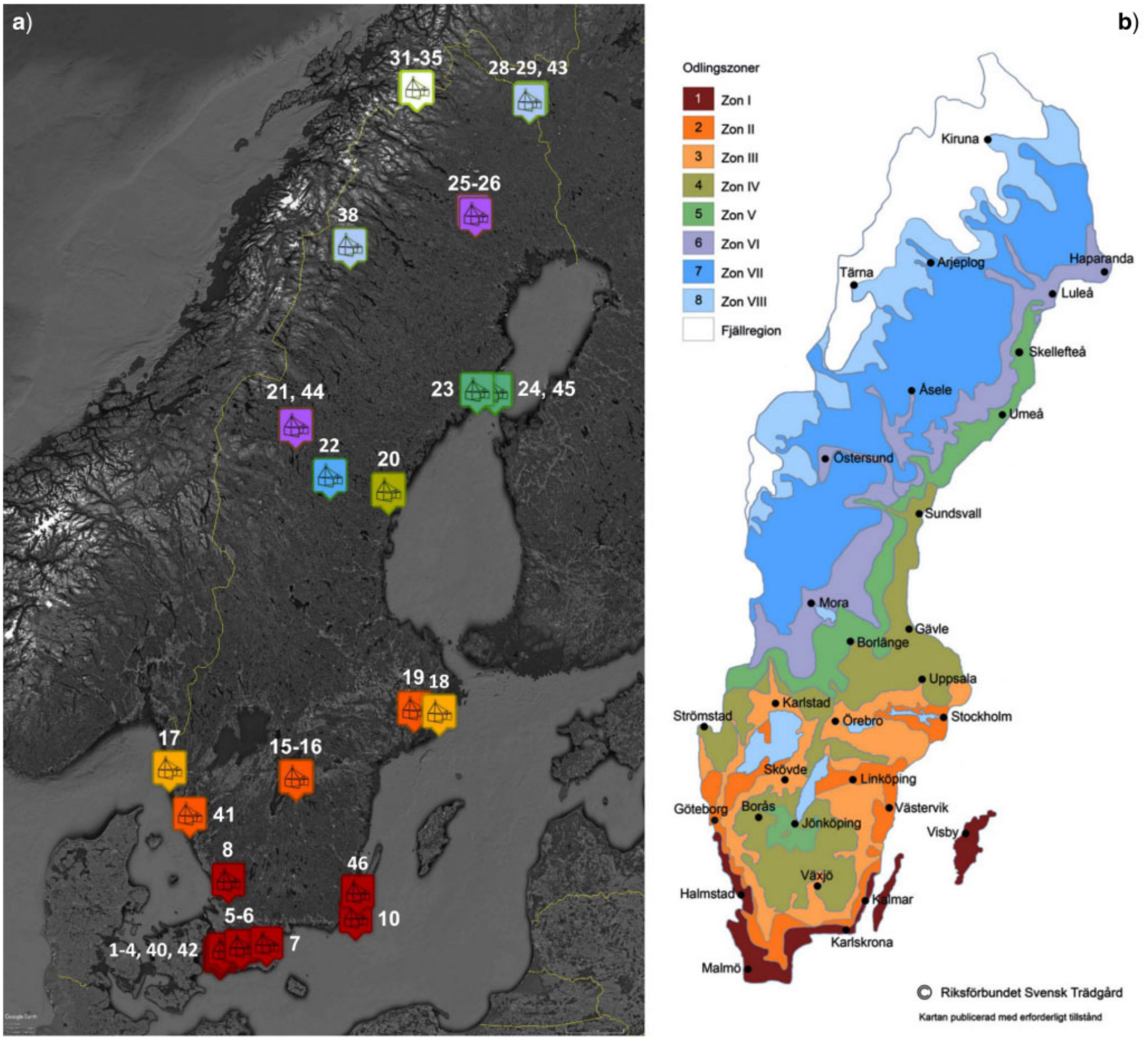

FIGURE 2. a) Sites of the Swedish Insect Inventory Project, color-coded by climatic zones identified by the Swedish Horticultural Society, b) Climatic zones (odlingszoner) of the Swedish Horticultural Society (Riksförbundet Svensk Trädgård), used with permission.

(fast basecalling model was used because the capacity for high accuracy basecalling was not available at this time).

### *Bioinformatics*

Demultiplexing and read filtering followed the protocols described in Wang et al. (2018) for Illumina and in Srivathsan et al. (2019) for MinION data. Illumina data processing involved the merging of paired-end reads using PEAR (v 0.9.6) (Zhang et al. 2014). Subsequently, reads were demultiplexed based on unique combinations of tags by an in-house pipeline that looks for perfect tags while allowing for 2-bp mismatches in primer sequence. For each sample, we retained reads that were longer than 300-bp after primers were removed and the identical reads (reads that only vary in length with terminal bases missing or with extra terminal bases) were merged and counted to identify the dominant sequence, that is, the sequence with the highest count in the data set. Barcodes were called when 1) the sample contained ≥50 reads and 2) the dominant read had a coverage of over 10×. Lastly, 3) these reads were accepted as the barcode for the specimen if the dominant sequence was at least 5× more common than the sequence with the next-highest abundance. For MinION data, miniBarcoder (Srivathsan et al. 2018, 2019) was used to demultiplex the data. Primers were found using *glsearch36*, and tags extracted allowing for 2 bp errors. Reads with a given tag combination were added into "specimen bins" and those bins having more than 5 reads were processed using multiple sequence alignment via MAFFT. Afterwards, a majority rule consensus barcode was obtained but only accepted if it had <1% *N* (or 4 *N*s in case of 313 bp barcodes). A second set of barcodes was obtained from the same bins using the consensus polishing tool RACON, where reads are mapped back onto the original MAFFT consensus barcode using GraphMap and RACON is used to call the consensus sequences (Sović et al. 2016; Vaser et al. 2017). The two sets of barcodes for the same reads were then corrected using amino acid translations that allow for resolving indel errors. Lastly, the AA-corrected sets of barcodes were consolidated as described in Srivathsan et al. (2019). This pipeline has been shown to provide >99.99% accurate DNA barcodes when compared to Illumina results (Srivathsan et al. 2019) and is available at https://github.com/asrivathsan/miniBarcoder.

### *Clustering and mOTU Estimation*

The DNA barcodes were aligned using MAFFT v 7.310 (Katoh and Standley 2013). Many different clustering algorithms for DNA barcodes exist, and we here initially used objective clustering (OC, Meier et al. 2006) but later compared the results with other methods. OC uses an *a priori* distance threshold to group sequences; cluster members are separated by at least this distance from members of all other clusters, but the maximum distance within a cluster can exceed the clustering threshold (Meier et al. 2006). As OC works based on a set threshold and is not an adaptive method, it is computationally inexpensive and allows for a straightforward comparison of criteria across clusters. Initial distance-based mOTU delimitation via OC used a 3% minimum pairwise interspecific threshold, which had previously been determined appropriate for phorids based on publicly available data (see Srivathsan et al. 2019). The clustering was achieved with a new implementation of the clustering algorithm implemented in SpeciesIdentifier (available at https://github.com/Gaurav/taxondna https://github.com/asrivathsan/obj_cluster). Cluster numbers referenced throughout the manuscript refer to these original 3% OC clusters. The clustering results were used to physically sort specimens into putative species in preparation for acquiring morphology as a second source of data. To visualize the barcode data for each cluster, we prepared median-joining haplotype networks with PopART (Leigh and Bryant 2015). These networks provided a good overview of the number, abundance, and distribution of haplotypes. The networks were color coded for geographic context based on the planting regions ("odlingszoner") recognized by the Swedish Horticultural Society (Fig. 2b, "Zonkarta för odlingzoner," used with permission) (Riksförbundet Svensk Trädgård 2018). These provide a breakdown of Sweden into nine zones that coincide with important shifts in the composition of the natural flora of the country based on climatic, geologic, and historical data. Using the odlingszoner offers a finer-scaled geographic approach than using broader classification schemes like forest type or climate zones.

### *Acquisition of Morphological Data for Barcode Clusters*

Morphological examination of specimens was first conducted in ethanol before some specimens were dissected and slide-mounted in Canada balsam following standard procedures for Phoridae (Disney 2009). Morphological examination relied upon a standard suite of characters and character states described for *Megaselia* (Hartop and Brown 2014; Hartop et al. 2016) and slightly modified for other genera. Morphological validation largely relied not only on characteristics of the male genitalia but also involved everything from overall gestalt, to setation of the thorax (especially of the notopleuron, scutellum, and anepisternum), legs, and frons, to characteristics of the wings. As LIT is designed to be efficient, morphological validation did not include more time-consuming morphological methods such as wing measurements or dissection of the genitalia.

### *Integration of Barcodes and Morphology I: Identifying Predictors of Incongruence*

We randomly selected 100 clusters from the 315 total using the RANDARRAY function in Excel (=RANDARRAY(100,1,1,315,TRUE)) for in-depth study of congruence between barcode and morphological data. The following rules were applied when selecting specimens for morphological study: i) Study at least one male (the sex used for species-level identifications for most Holarctic phorids) for all main haplotypes (= those containing 20% or more of the specimens), ii) study additional specimens within the 3% cluster until no haplotypes remain that are >1 bp away from any checked haplotype, and iii) pick at least one specimen from each horticultural zone represented in the cluster. In most cases, following these procedures was straightforward. For example, Cluster 293 (Fig. 3) has two main haplotypes that each contains >20% of specimens and a singleton that is 2 bp away from the checked haplotypes, meaning a minimum of three specimens had to be checked according to the haplotype rules. However, the cluster also occurs in seven zones so that a specimen for each zone had to be checked raising the total number of specimens to at least seven specimens to accommodate the zone and haplotype rules. The selection of specimens can be ambiguous because multiple sets of specimens satisfy the stipulated criteria. In these cases, we arbitrarily chose one of the sets. In the case of a rejected initial hypothesis (cluster containing multiple lumped morphospecies), additional specimens had to be examined to determine which haplotypes within the barcode cluster belong to which morphospecies. In most cases, this was straightforward: intermediate haplotypes were checked guided by haplotype networks. In one

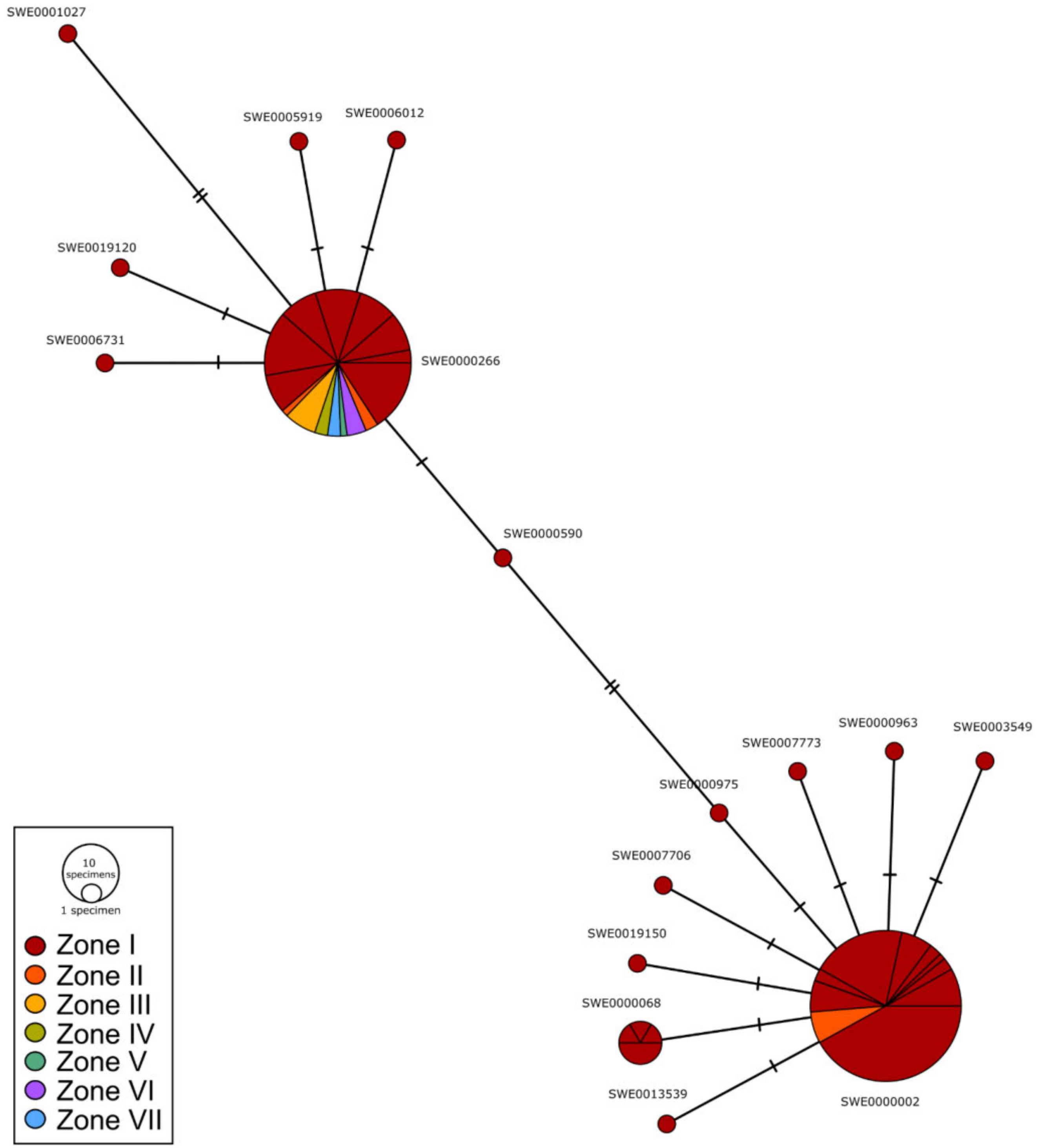

FIGURE 3. Haplotype network for Cluster 293, color-coded according to the climatic zones of the Swedish Horticultural Society. Nodes represent each unique haplotype, pie slices of nodes indicate the proportion of specimens from a particular site, node diameters are proportional to the number of specimens the haplotype contains, and the lines connecting the nodes have hash marks corresponding to base pair differences.

case, however, a single barcode cluster contained at least 25 morphospecies (Cluster 101: Fig. 4) and had to be reclustered at successively lower thresholds (2% and 1%) to help with specimen selection and species delimitation. After reclustering at lower thresholds, the standard checking rules were applied to the subclusters.

Note that for all clusters, only males were considered fully informative as temperate zones are dominated by species that require males for morphological

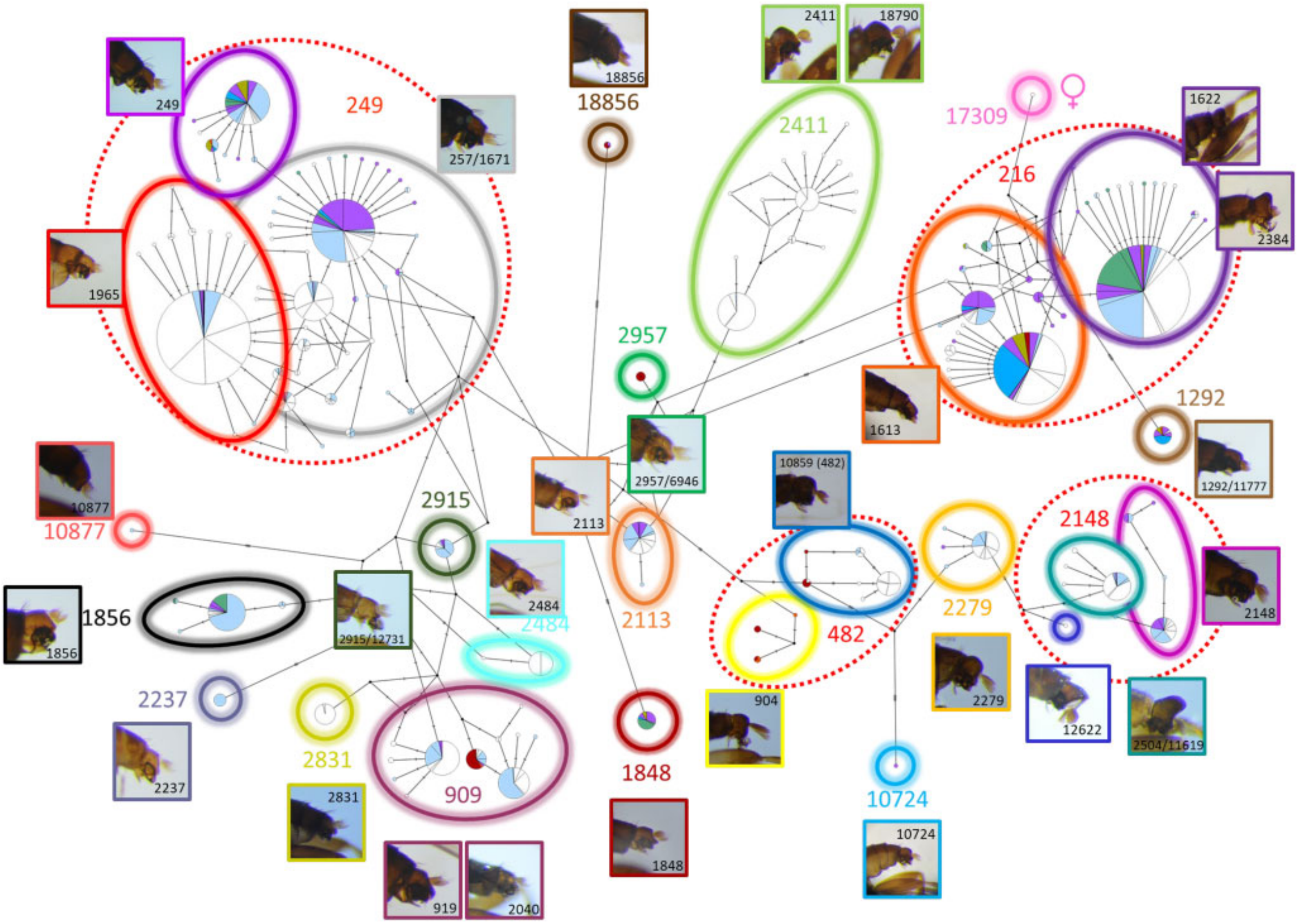

FIGURE 4. Haplotype network for cluster 101 indicating all morphological species found with male genitalia illustrated (border colors of genitalia figures match morphospecies boundary colors). Morphospecies is equivalent to 1% clusters (indicated by numbers), except in cases where a 1% subcluster contained multiple morphospecies, in these cases the 1% cluster is a red dashed line around the morphospecies. For two subclusters (216 and 249), the network is too complex to accurately circumscribe morphospecies in this figure. Morphospecies designations for all specimens are in the cluster table available on the project GitHub page.

identification. Key diagnostic characters of females examined were always congruent with the males for the cluster, but they were not considered sufficiently informative to evaluate distant haplotypes (>1% or >4 bp). A single exception to this was a distant female haplotype in a cluster of *Spinophora* that could be reliably identified to species. Nonsingleton clusters that only contained female specimens, or distant haplotypes represented only by females were removed from the analysis as they could not be properly validated.

To test which cluster-specific properties of 3% clusters were most effective at predicting incongruence with morphology, we fitted a generalized linear model (glm) with quasibinomial errors to the data obtained for the 100 randomly selected clusters. The response variable was "validated" (if a cluster was congruent with morphology) and the explanatory variables were six cluster properties: "haplo" (number of haplotypes in a cluster), "spec" (number of specimens in a cluster), "stability" (see below), "max_p" (maximum pairwise distance within a cluster), "zones" (number of geographic zones represented in a cluster), and "sites" (number of sites represented within a cluster).

*Cluster stability* quantifies whether a barcode cluster is sensitive to changes in clustering thresholds. This is formalized as Class II specimen congruence in Yeo et al. (2020). For a set of clusters at 1%, $A_{1a}, A_{1b}, A_{1c} \ldots A_{1n}$, that combine to form a single 3% cluster, $A_3$, the stability value is given by $\frac{max(|A_{1a}|,|A_{1b}|,|A_{1c}|...,|A_{1n}|)}{|A_3|}$, where $A$ is the set of unique haplotypes and $|A|$ is the number of elements in $A$. Simply, this is the number of haplotypes contained in the largest 1% cluster that is found within a 3% cluster, divided by the number of haplotypes in that 3% cluster.

A correlation matrix was used to examine collinearity between explanatory variables (corrplot.mixed in R package corrplot) (Supplementary Fig. S1) before using the Farrar–Glauber test (R package mctest) to detect and remove variables systematically according to the variable inflation factor, rerunning the model until collinearity was no longer detected and the remaining variable(s) were statistically significant. Two clusters were designated "species complexes" (see Results section) because they had uncertain species numbers based on morphological evidence. We thus constructed the model twice: once with these counting as validated clusters corresponding to a single morphospecies, and once with them counting as incongruent (containing multiple species).

### *Integration of Barcodes and Morphology II: Validating Predictors of Multispecies Clusters*

After developing predictors of incongruence based on the first 100 clusters, the remaining 215 were used to test the two most important predictors of incongruence: low cluster stability (<1.0) and large intracluster p-distances (≥1.5%). Forty-three of the 215 clusters were identified as potentially incongruent (PI) while 172 were non-PI. The 43 largest non-PI were used as control for the 43 PI clusters. These 86 test clusters went through the following validation process: 1) We checked one specimen each for all main (>20% of specimens) haplotypes and 2) a pair of specimens representing the maximum p-distance in the cluster. For the selection of most distant haplotypes, we ignored small sequence differences (1–2 base pairs) in favor of sampling main haplotypes (see Results section). Using our previous example of Cluster 293 (Fig. 3), the two main haplotypes would be checked, but the haplotypes "haloing" the main haplotypes by 1–2 base pairs would be ignored. If a cluster was found to contain multiple morphospecies during the initial check, additional specimens were studied across the cluster to delimit morphospecies boundaries. For the remaining 129 non-PI clusters, we only checked a pair of specimens representing the maximum p-distance in the cluster.

### *Evaluation of the Performance of Different Clustering Algorithms and Thresholds*

We varied the threshold for OC and evaluated the performance of three alternative procedures for barcode clustering. The first was Automatic Barcode Gap Discovery (ABGD) (Puillandre et al. 2012), the second Poisson Tree Processes (PTP) (Zhang et al. 2013), and the third Refined Single Linkage (RESL) (Ratnasingham and Hebert 2013). ABGD is distance-based like OC and attempts to find a barcode gap (difference between inter- and intraspecific distances) for each subgroup of sequences based on an iterative application of priors for intraspecific divergence and identification of the first significant gap beyond this divergence. PTP is tree-based and uses branch lengths to estimate transition points between intraspecific and interspecific branching in an input phylogeny, thereby determining which monophyletic groups likely consist of a single species. RESL is used to calculate Barcode Index Numbers (BINs) on the Barcode of Life Data System (BOLD). The underlying algorithm is not public. Instead, a new BIN classification is generated monthly by the Canadian Centre for DNA Barcoding.

OC was carried out at 0.0–5.0% uncorrected p-distance thresholds at 0.1% intervals (thresholds between 0.5% and 4.0% were considered for delimitation comparison, see Results section). For ABGD estimation of mOTUs, we used the default range of priors ($P = 0.0010$, $P = 0.0017$, $P = 0.0028$, $P = 0.0046$, $P = 0.0077$, $P = 0.0129$, $P = 0.0215$, $P = 0.0359$, $P = 0.0599$); these priors represent the maximum intraspecific divergence in the first iteration of the algorithm. The slope parameter (−X) was reduced in a stepwise manner (1.5, 1.0, 0.5, 0.1) if the algorithm could not find a partition, as done by Yeo et al. (2020). All other parameters were kept as default. For PTP, we used RAxML v8.4.2 (Stamatakis 2014) to estimate the topology under the GTRGAMMA model. Twenty independent searches were conducted, and the best scoring topology across these searches was retained. mOTUs were then obtained based on the application of the PTP model on this topology, as implemented in the mPTP software (–single –ml mode) (Kapli et al. 2017). We used two indirect methods for evaluating RESL performance for our data, as the algorithm is not public (Meier et al. 2021), and the current implementation of the algorithm does not assign new BINs to minibarcodes. We initially identified those haplotypes in our data set that have a 100% match to haplotypes that are already on BOLD and assigned the corresponding BIN numbers to the barcodes for our specimens with these haplotypes. Afterwards, two congruence analyses were carried out. One was based on the clustering using all our data, but only scoring congruence for specimens with BIN numbers. A second was based on a reanalysis (reclustering with OC) of only those haplotypes with BIN numbers. Note that both comparisons could be impacted by differences in the underlying data that are used to assign barcodes to BINs.

Congruence between mOTUs and morphospecies was assessed based on match ratio as described in Ahrens et al. (2016). The match ratio is computed as $2 * N_{\text{match}}/(N_x + N_{morph})$, where $N_{\text{match}}$ is the number of completely matching clusters, $N_x$ is the total number of clusters (mOTUs) identified by method $x$, and $N_{\text{morph}}$ is the total number of morphospecies. We also evaluated the performance of clustering methods by determining the number of mOTUs that contained several morphospecies (merged clusters, $N_{\text{merged}}$), that contained only part of one morphospecies (split clusters, $N_{\text{split}}$), or the few cases where a method both split and merged members of a single morphospecies $N_{\text{split/merged}}$. Note that $N_x = N_{\text{match}} + N_{\text{merged}} + N_{\text{split+}} N_{\text{split/merged}}$. Next, we recorded the total number of specimens in each type of cluster. The results were visualized with nVenn (Pérez-Silva et al. 2018) for both optimal (best match to morphospecies) and conservative (above splitting of morphospecies) settings.

To formalize our methods, we created a pairwise distance matrix with MEGA X with default settings (Kumar et al. 2018). This was used as input in the creation of a specimen-picking algorithm (Supplementary Fig. S2). Using this algorithm, we computed the minimum number of specimens that must be examined morphologically to validate barcode clusters for different methods and settings if the following simplified sampling scheme was applied: 1) determine if the cluster is PI or non-PI (here, based on maximum p-distance only, see Discussion section), 2) check the main haplotypes and two most distant haplotypes for PI clusters, and two most distant haplotypes only for non-PI clusters. If checked specimens do not belong to the same species, 3) check additional haplotypes until no unchecked haplotypes remain that differ by >1 bp from any checked haplotype. Additionally, all checked haplotypes separated by this distance must belong to the same morphospecies (i.e., there are no unchecked specimens between specimens that belong to two different morphospecies). We also calculated the number of morphospecies (if any) that would be overlooked by this minimal sampling procedure.

### *Data and Code Availability*

Data and scripts are available at https://github.com/ronquistlab/taxon-cluster-paper.

## Results

### *Barcoding and Initial Clustering*

The 19,570 phorids in the sample yielded 17,902 barcodes in 1,714 haplotypes and 340 clusters at a 3% OC threshold. A BLAST search revealed 11 clusters with best matches >95% to nonphorid taxa. These were examined, and six were removed from the data set: five were confirmed to be specimens from different Diptera families (incorrectly sorted during preprocessing: 9 clusters totaling 10 specimens), and a singleton cluster was found to lack a voucher specimen because it was lost during molecular processing. The remaining clusters were confirmed to be phorids, likely these were matched to database misidentifications (a common problem in dark taxa). During morphological evaluation of clusters, four additional clusters were found to be other Dipteran families and removed, and one cluster was found to be contaminated (a single cluster of 42 phorid specimens that contained a mix of disparate species). Finally, of the 329 clusters, 14 clusters (406 specimens) could not be rigorously evaluated with morphology because they either contained only female specimens, or the most-distant haplotypes were only represented by females. The final analysis thus included 17,443 specimens in 315 clusters obtained at a 3% OC threshold.

Of the 315 3% clusters that were morphologically evaluated, 5.7% (18/315) contained multiple morphospecies. All morphospecies were congruent with haplotype clusters delimited at OC thresholds lower than the original 3%; that is, the morphospecies formed disjoint haplotype subsets of the original 3% haplotype set. This means that morphospecies were congruent with the barcode data clustered at alternative thresholds and only incongruent with barcode clusters obtained at 3%. These 18 clusters contained 18.6% of the morphospecies (68/365), but were also among the most abundant in terms of specimen numbers. For example, 19.6% of all specimens were found in the largest incongruent cluster (Cluster 101) while an additional 8.8% of specimens belonged to the second largest cluster (Cluster 68). A total of 7,150 specimens (41% of the total) belonged to clusters that contained more than one morphospecies. All multimorphospecies clusters belonged to the genus *Megaselia*. There were no cases of morphospecies consisting of haplotypes from multiple 3% OC clusters.

### *Integration of Barcodes and Morphology I: Identifying Predictors of Incongruence*

Of the 100 3% clusters randomly selected for the exploratory phase, seven contained multiple morphospecies and two clusters were found to contain morphological variation that was suggestive of species complexes that required more data for resolution. Of the seven incongruent clusters, six split into two morphospecies each, while one cluster (Cluster 101: 3421 specimens) contained at least 25 morphospecies (exact species count will require more data to resolve some of the subclusters) (Fig. 4). The 7% multimorphospecies clusters (7/100) contained 28% of the morphospecies (37/130) and 58.6% of the specimens (3501/5977).

Although the validation procedure involved the examination of many specimens, we noted that the multispecies nature of a cluster was always revealed by the morphological differences between specimens representing the two most distant haplotypes. The requirement to examine specimens from all horticultural zones never led to the discovery of additional species. Similarly, we found no additional species in the "satellite haplotypes" that often "halo" around main haplotypes and differ only by 1–2 bp (see Fig. 3). We therefore removed checking specimens for these as requirements in subsequent stages.

The exploratory study suggested that the number of specimens ("spec"), the number of haplotypes ("haplo"), the number of collecting sites ("sites"), the number of horticultural zones ("zones"), the maximum p-distance ("max_p") and the stability to varying clustering thresholds ("stability") might be correlated with the presence of multiple morphospecies within clusters, and all of these were initially included in our model. Three pairs of variables had high collinearity: "max_p" and "stability" (0.77), "haplo" and "spec" (0.99), and "zones" and "sites" (0.79). We systematically removed variables from the model based on the variable inflation factor until collinearity was no longer detected and were left with "spec," "stability," and "zones," but only "stability" was significant (Supplementary Fig. S1). Results of the linear model were the same regardless of whether the two species complexes were classified as validated or not (Supplementary Fig. S1).

Our analysis therefore indicates that only the variable "stability" is a significant predictor of cluster failure, but the high collinearity (0.77) between "max_p" and "stability" implies that it, too, would be predictive (Supplementary Fig. S1). We subsequently noticed that only six of the seven multimorphospecies clusters were unstable but all of them had high p-distance. This suggested that p-distance should also be used to identify PI clusters. Using only the p-distance or the two variables in combination (either/or) ensured that all seven clusters were identified but it also significantly increased the false positive rate: 22 clusters had high maximum p-distance, but just 12 of these were unstable. Although the combination of the two variables yields a high false-positive rate, we used both criteria for identifying PI clusters among the remaining 215 clusters. More efficient processing could be achieved by dropping one criterion. However, this would result in overlooking the incongruence between data sources for some clusters.

### *Integration of Barcodes and Morphology II: Validating Remaining Clusters and Testing Predictors for Multispecies mOTUs*

After an in-depth study of the first 100 clusters, 215 OC clusters remained (3% threshold). We identified 43 clusters that had signatures of incongruence because of high maximum distances (>1.5%: 14 clusters), instability (stability < 1.0: two clusters), or both (27 clusters). Of these PI clusters, a large proportion (26%: 11/43) containing 49% of the species (31/63) failed morphological validation, and one additional cluster was classified as a "species complex." All 11 failed clusters had both high maximum *p*-distances and instability, while the species complex was identified by maximum p-distance only. To determine whether 26% failure is an unexpectedly large proportion of failing clusters, we used the largest 43 non-PI clusters as control. The largest non-PI clusters were used to best match cluster sizes for PI (mean = 152 specimens) and non-PI clusters (mean = 95). All 43 non-PI clusters passed morphological validation as did the remaining 129 smaller, non-PI clusters.

The total number of specimens examined for this study was 1,039, or 5.8% of the total. This includes the specimens for the more extensively sampled first 100 clusters. It also includes some female specimens that represented haplotypes for which a male was also examined. Without these additional specimens, the number of specimens needed for validation of 315 3% OC clusters would have been 915 (5.1% of total). The optimized number of specimens that needs to be studied is 861 (4.8% of the total) when using OC 1.3–1.5% or ABGD 0.0077 as clustering thresholds. This is an average of 2.3 specimens per species.

### *Impact of Barcode Clustering Algorithms on Results*

We clustered across a wide range of thresholds (0–5.0% OC) to assess the structure of variation in our data set. We would not consider thresholds lower than 0.5 or above 4.0% to be appropriate for initial delimitations in most groups unless there is evidence for unusually low or high genetic variation within species. Narrowing the range of thresholds accordingly, the number of clusters across methods and settings varies from 424 (0.6% OC) to 207 ($P = 0.0359$ ABGD) (Fig. 5). Our morphological study suggested the presence of 365 morphospecies and a match ratio of 0.871 for 3% OC clusters (Fig. 6, Supplementary Table S2). All incongruence between 3% OC clusters and morphology was due to lumping. Lowering the threshold to 1.6–1.7% maximized the match ratio for OC at 0.897 (Fig. 6, Supplementary Table S2). ABGD's highest congruence was the same (0.897) and observed when the prior of intraspecific divergence (p) was set to 0.0077 (Fig. 6, Supplementary Table S2). PTP fared less well, with a match ratio of 0.841 (Fig. 6, Supplementary Table S2). Regarding RESL, 277 of the 1714 haplotypes in our data set had 100% matches to publicly available sequences with BIN designations. These 277 haplotypes represented 50% (186) of our morphospecies in 172 BINs. Of the 186 morphospecies matched to BINs, 162 (86%) were congruent with BINs. These 277 haplotypes were also subjected to reclustering with several algorithms. We compared the results for both optimal (highest match ratio) and conservative (above cluster splitting) threshold settings for OC and ABGD. BIN designations again matched 86% of morphospecies. This was better than OC at the conservative threshold of 3% (76% of morphospecies correctly delimited) but worse than OC at the optimal 1.7% setting (89% of morphospecies correctly delimited). PTP performed poorer for this reduced subset (74%), and ABGD had the lowest level of congruence (correctly delimiting 72% of morphospecies at $P = 0.0077$ but only 51% at $P = 0.0215$).

We then examined whether morphospecies were split, lumped, or split+lumped across different methods and settings (Fig. 7, Supplementary Table S2). This revealed that OC starts lumping morphospecies at 0.6% and stops splitting morphospecies at 2.8%, while ABGD lumps morphospecies across all priors, and stops splitting at $P = 0.0215$ (Fig. 7, Supplementary Table S2). PTP both splits and lumps morphospecies (Supplementary Table S2). There were very few cases where a method both split and lumped a single morphospecies, but this was the case for one morphospecies using PTP and at OC thresholds 0.7–0.9% (Fig. 7, Supplementary Table S2). RESL split two of the clusters designated as "species complexes" in our analysis but lumped 22 morphospecies into BINs (from OC clusters 68, 79, 91, 101, and 103). The BIN composition of the complex Cluster 101 could be assessed for 16 morphospecies (Fig. 8). Based on this partial representation, RESL lumps eight of the morphospecies from Cluster 101 into a single BIN (BOLD-AAG3235, shown in red) and another three into a second BIN (BOLD-AAL9067).

Figure 9 illustrates congruence between methods for optimal (OC 1.7, ABGD $P = 0.0077$) and conservative settings (OC 3.0%, ABGD $P = 0.0215$). At optimal settings, 313 clusters were the same across methods and ABGD and OC results were 100% congruent. PTP

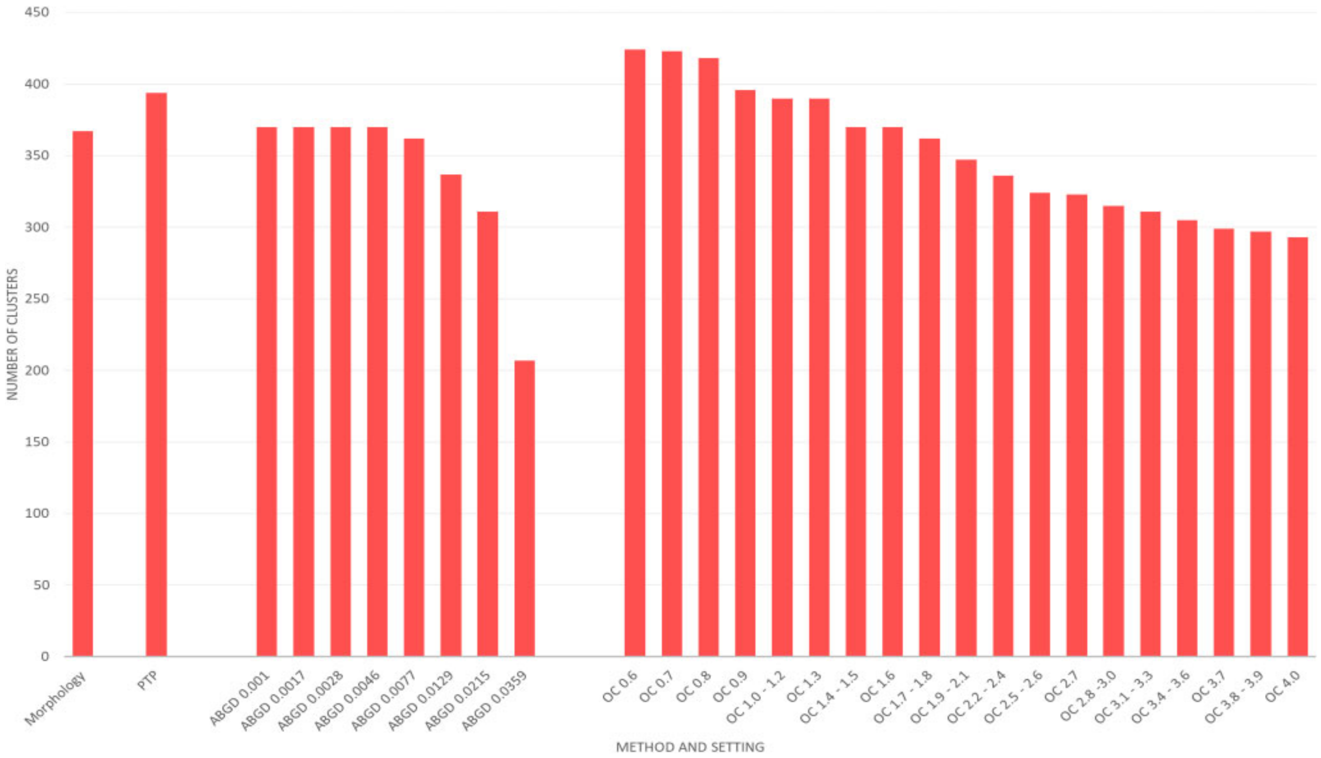

FIGURE 5. The number of morphospecies, and clusters across settings with PTP, ABGD, and OC. OC is plotted without 0–0.5% thresholds where 1–2 bp differences between haplotypes greatly inflated cluster numbers.

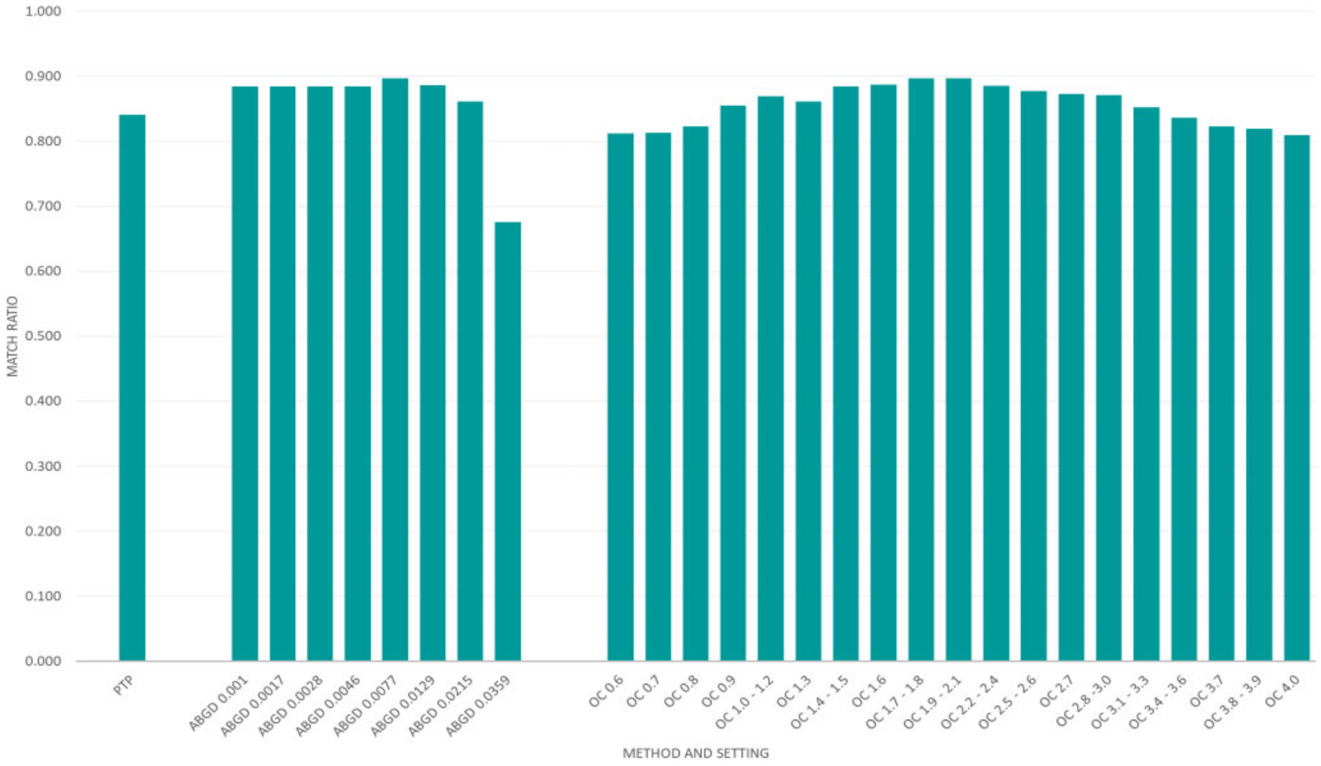

FIGURE 6. Match ratios for PTP, ABGD (all priors), and OC (all thresholds) versus morphology across methods and settings.

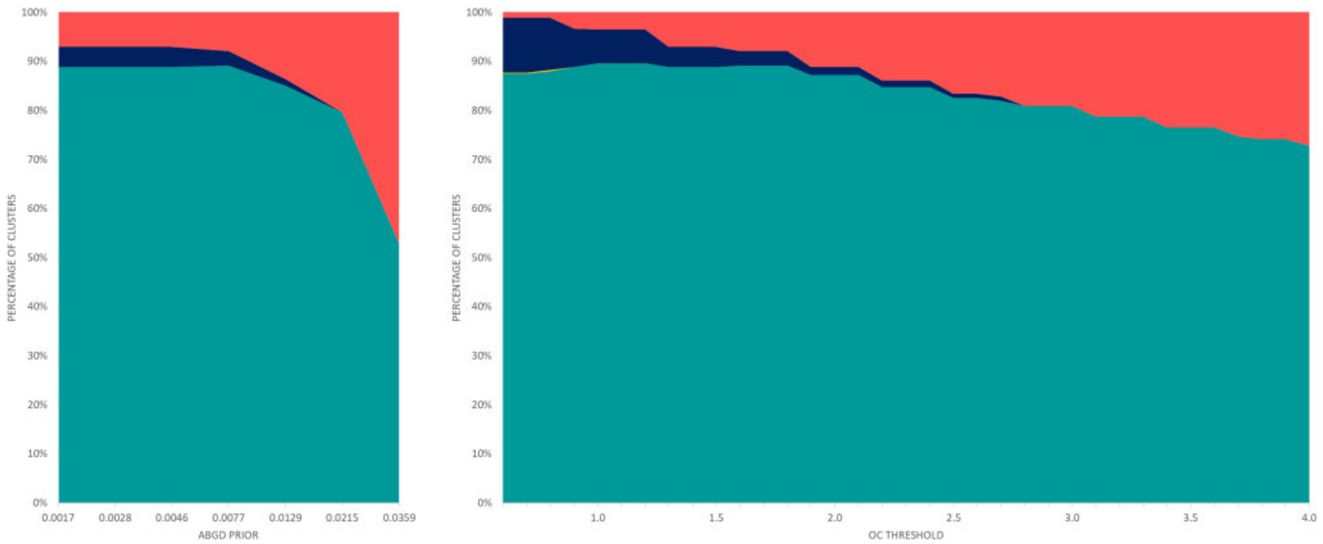

FIGURE 7. The correct delimitation (teal), splitting (dark blue), lumping (coral) and splitting/lumping (yellow) of morphological clusters with ABGD (left) and OC (right) across settings. A color version of this figure appears in the online version of this article.

was an outlier, as it tended to split species compared to the other methods. Morphology suggested several dozen outlier species that were different from what was inferred with all molecular methods. This was due to multiple morphospecies lumped by ABGD and OC at conservative settings (including the 25 species in Cluster 101), while at optimal settings the difference was due to fewer lumped morphospecies, but additional split morphospecies.

## DISCUSSION

The goal of LIT is to transform the study of dark taxa. These hyperdiverse groups are currently neglected because traditional methods are unsuitable for samples containing thousands of specimens and hundreds of species. LIT takes on the challenges that these taxa pose by overcoming the conundrum of how biodiversity discovery based on multiple data sources can be accelerated when most traditional studies only use one and are still too slow. We agree with Puillandre et al. (2012) that this can be achieved by starting with one type of data that can be acquired rapidly and inexpensively, but it would be too slow and expensive to acquire a second or even a third data source for all specimens. Instead, we find that there are systematic ways to flag those barcode clusters and specimens that are most likely to be incongruent with morphospecies boundaries. This allows for the second type of data to be acquired for only a few specimens, thus saving time and expenditure while allowing for integrative species delimitation in dark taxa. Critically, LIT has the potential to require little expert time if the primary data source (e.g., NGS barcodes or images) is obtained by nonspecialists as a first step in sample processing. This is a realistic scenario because primary data sources like barcodes are suitable for automation. Scientists will then work only on material presorted to putative species, allowing them to focus on a small number of specimens from clusters that have a high probability of being incongruent with secondary data sources (e.g., morphological or nuclear data). LIT ensures that all species are covered by at least two types of data that can be summarized automatically in preparation for (re)descriptions based on established and efficient methods (Butcher et al. 2012; Riedel et al. 2013; Lücking et al. 2016). An additional benefit to LIT is that it is likely less costly than many recently funded collection digitization initiatives because most label information for incoming samples will already be digitized and even the imaging of specimens can be automated if devices such as the DiversityScanner are used (Wührl et al. 2021).

### *LIT of Swedish Phorids: Clustering with the First Data Source*

We used NGS barcodes as our first data source because they are readily acquired for phorids and suitable for creating preliminary species hypotheses through clustering. There are many algorithms for clustering barcodes. Most methods include the disclaimer that they should not be used to delimit species without consulting other evidence (Puillandre et al. 2012; Ratnasingham and Hebert 2013; Zhang et al. 2013), but it is not uncommon to see molecular clusters equated to species without, or with an unexplained process of validation. Our study confirms that none of the sequence clustering methods tested (OC, ABGD, PTP, and RESL) accurately delimit morphospecies across taxa with disparate evolutionary processes, thus confirming the need for integrative methods. Adaptive algorithms, such as ABGD and RESL, should be able to accommodate biological variation across subgroups better than methods based on fixed thresholds, but our results indicate that a simple method using fixed thresholds (OC) does as well or better than

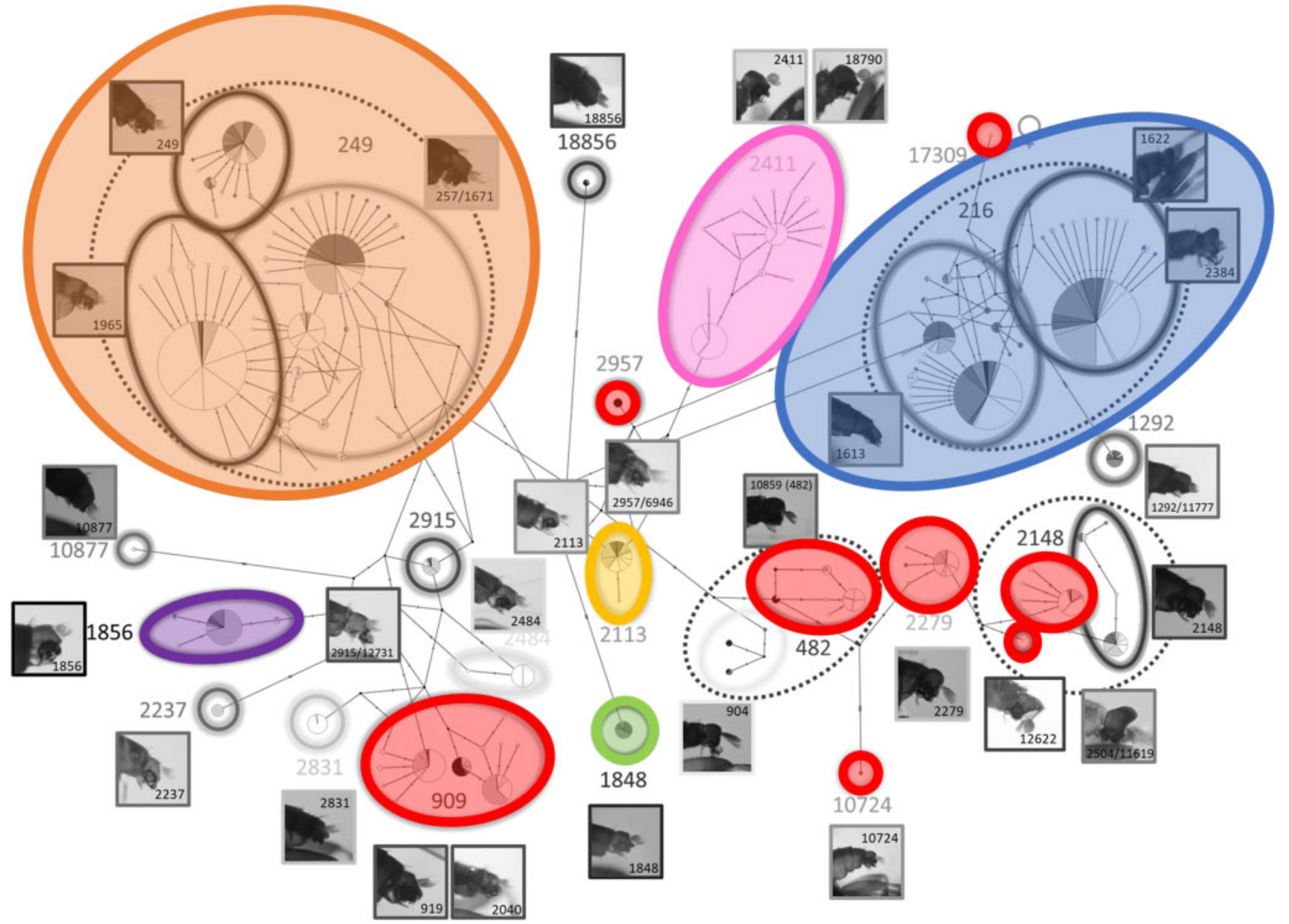

FIGURE 8. BIN designations (each BIN designated by a different colour) of the 16 morphospecies of Cluster 101 for which we found a 100% match (to at least one specimen) in BOLD.

adaptive methods. The only tree-based method (PTP) did the worst of the tested methods, possibly because the tree reconstruction was based on minibarcodes that provided very limited information for reconstructing phylogenetic relationships and identifying the boundaries between within-species and between-species tree structures. Tree-based methods should be tested with trees that are reconstructed based on multiple markers. Such data would also be required for other methods that use multispecies coalescent and related models. LIT will also be relevant for these methods because it predicts for which specimens/haplotypes multiple markers need to be collected for rigorous species delimitation. Fortunately, our study suggests that most species can be delimited using minimal amounts of data and only species complexes require more markers for resolution.

### *Incongruence and Conflict between and within Data Sources*

Overall, the congruence between molecular clusters obtained with different methods was better than with morphospecies. For example, at optimal thresholds, OC and ABGD produced the same clusters across the data set, but an additional 40 morphospecies were found based on morphology (Fig. 9a). Similarly, for the data used to evaluate RESL, morphology suggested the presence of 14 species that were not delimited by any of the molecular methods (Fig. 9c). Should we take these cases as evidence that the morphological evidence is misleading? This would be perilous given what is known about barcodes and the algorithms that are used to cluster them. Even the best algorithms will not be able to accurately delimit species if speciation has left no trace in the COI data, as is expected for recently evolved species. For example, PTP was introduced with the "fundamental assumption…that the number of substitutions between species is significantly higher than the number of substitutions within species" (Zhang et al. 2013). Similarly, with ABGD, "the lower the speciation rate, the better the performance of the method" (Puillandre et al. 2012) and RESL carries the warning "closely related species…will be overlooked because of their low sequence divergence" (Ratnasingham and Hebert 2013). In addition, dense specimen sampling may reveal near-continuous sequence variance across thousands of specimens (see Cluster 101) so that any algorithm will struggle to find species.

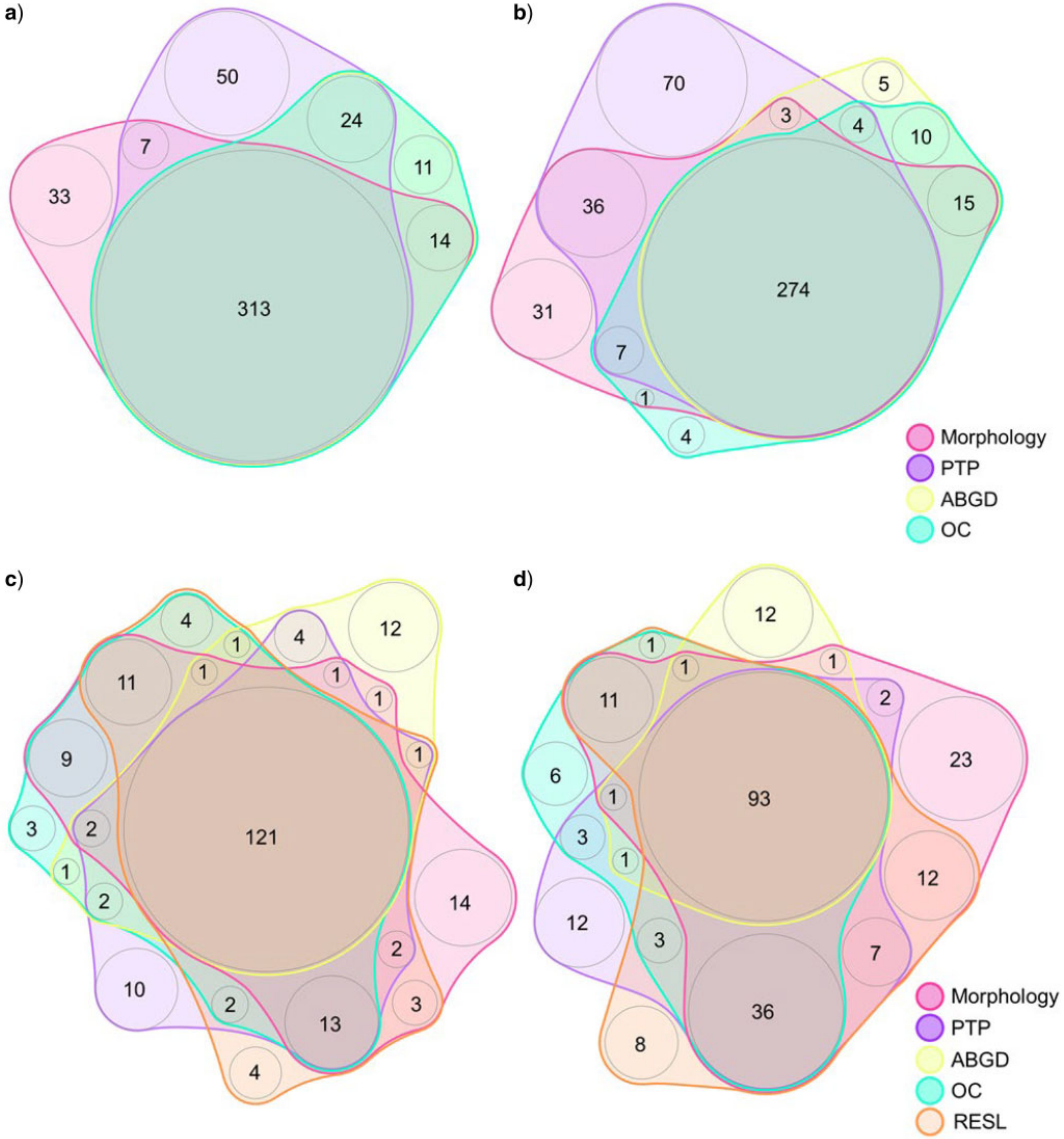

FIGURE 9. Congruence between morphology, PTP, ABGD, and OC methods with a) optimal settings (ABGD $P=0.0077$, OC 1.7%) and b) conservative settings (ABGD $P=0.0215$, OC 3.0%) and between morphology, PTP, ABGD, OC, and RESL methods with c) optimal settings (ABGD $P=0.0077$, OC 1.7%) and d) conservative settings (ABGD $P=0.0215$, OC 3.0%).

Overall, it is expected that barcode clusters will be incongruent with morphology for recently diverged species which will be lumped into one barcode cluster. This is different from what we would call "genuine conflict" between morphology and barcode data. Such conflict would be present when a morphospecies is composed of a collection of haplotypes that is inconsistent with all ways to cluster barcode data with an algorithm or threshold that was selected *a priori*. If such cases were observed, one would need to use at least one additional data source to test whether morphology and/or barcodes yielded misleading grouping statements. These additional data could be life history, ecological data, and/or nuclear markers (Puillandre et al. 2012). The standardization of the latter will be particularly important in the future (Eberle et al. 2020), as it will allow institutions to cover dark taxa for which the world lacks taxonomic expertise by using barcodes for presorting and nuclear markers for validation (or vice versa).

In our study, we found that of the 315 initial 3% clusters, most (297) were congruent with morphology and thus likely represented species. The remaining 18 contained multiple morphospecies. However, all morphospecies formed disjoint haplotype subsets of the 3% clusters. Given that these subsets can be obtained by clustering the haplotypes at lower thresholds, the morphospecies was "only" in conflict with a specific way to cluster the barcode data (e.g., a specific clustering threshold), but not with other grouping statements that could be supported by the barcode data. Which threshold is overall most appropriate for generating an initial set of species hypotheses is based on *a priori* estimates. In our study, we used 3% OC, because we had previously found that this threshold maximized congruence between phorid species names and barcodes in Genbank (Srivathsan et al. 2019). However, a single threshold is unlikely to be appropriate for all species, and even adaptive algorithms cannot separate species that have identical, or nearly identical, haplotypes. Therefore, we used a second source of data (e.g., morphology) to first test whether standard barcode clusters were species and when this was not the case, we determined whether there is a hierarchical level at which a barcode cluster is congruent with a morphospecies. All incongruence disappears once barcode data are allowed to support subclusters as species. For example, 3% OC clusters are generally composed of several disjoint cluster subset(s) that can be obtained by clustering the haplotypes at lower thresholds (e.g., 1% or 2%). Overall, this observation is consistent with the expectation that barcodes will lump closely related species.

This raises the question whether the groupings made congruent through threshold adjustments should be accepted as species or need validation with a third data source. We would submit that there is no clear-cut answer. Demanding the use of additional data will slow down taxonomic progress and may do more harm than good given that lumping recently diverged species is expected for barcode data. We would, however, propose that a third type of data should be obtained when the clustering thresholds have to be strongly modified compared to what is normally observed for a taxon. For example, in Cluster 101 there are morphospecies that are separated by just 0.6%. Since this is an unusually small separation between species, far below the initial clustering threshold, a third data set is needed for resolving these species. In other cases, very minor changes to the standard threshold yields congruent groupings, and we would submit that real progress for dark taxa can only be made if one were to go ahead and consider the evidence for these taxa to be sufficiently strong to treat them as species.

### *LIT of Swedish Phorids: Indicators of Potential Incongruence*

After initially clustering the barcodes of our specimens at 3%, we needed to determine what properties of such clusters increased the likelihood of incongruence with morphology as the validation data. We tested several properties and found that cluster stability is the best predictor for incongruence, but we also used maximum p-distance because it was shown to be collinear with stability and some incongruent clusters were only identified by this variable. Cluster instability is consistent with areas where multiple species are separated by shallow (below initial clustering thresholds) splits, while maximum p-distance identifies clusters with unusually high variation. Such shallow splits between species pose problems for species discovery with barcodes, but they are expected because few evolutionary biologists doubt that there are cases of recent and rapid speciation (Puillandre et al. 2012; Ratnasingham and Hebert 2013; Zhang et al. 2013). We here quantify cluster stability by testing whether cluster membership changes when the clustering threshold is modified. This is suitable for OC but an alternative way to identify unstable clusters would be to inspect the longest branch length on a median-joining network for each cluster (as in the haplotype networks) (see Supplementary Fig. S2 for how this might be incorporated into the specimen checking algorithm). High within-cluster distances may be indicative of two species or old lineages within one species that have acquired large amounts of genetic variation but have failed to speciate. Overall, it is thus not surprising that these cluster properties are associated with incongruence, but it was surprising to find that they are so effective at predicting incongruence. Over one-quarter of the 3% clusters flagged by the predictors failed validation with morphology, while none of the control, non-PI clusters contained more than one morphospecies. Of course, this result will have to be tested for more taxa and larger samples before a more general use of these predictors can be advocated.

### *LIT of Swedish Phorids: Selective Use of Second Data Source*

Identifying 3% of clusters that are likely in need of refinement is only the first step. LIT also needs rules for selecting those specimens that should be studied using a second type of data (e.g., nuclear genes and morphology). In this context, it is important for taxonomic work on dark taxa to minimize the number of specimens examined while ensuring that all clusters containing multiple species are reliably discovered. Our final LIT protocol for this study includes specimen-picking recommendations (Fig. 1) that are formalized as an algorithm available from the project's Github repository (Supplementary Fig. S2). The basic recommendation is simple and common sense, but our large-scale study shows that it is also effective. All nonsingleton clusters are designated as either PI or non-PI (although singleton clusters could be designated PI based on stability if examined at an increased, rather than decreased, threshold). PI clusters are then validated by sampling the main and most distant haplotypes, while the verification of non-PI clusters is based on a pair of specimens representing the most distant haplotypes. Based on our data clustered at 3%, this protocol will reliably distinguish clusters that are congruent with morphospecies from those that are incongruent. Resolving the incongruent clusters then requires more in-depth examination. For our data, we demonstrate that only ca. 5% of all specimens (i.e., 2.5 specimens per species; 915 specimens in total) need to be checked to integrate barcode data with morphology.

The LIT protocol used here includes checking main haplotypes for clusters suspected of incongruence because multiple morphospecies can be intermixed within a closely related network (e.g., "subcluster 216" in Cluster 101, Fig. 4). Checking specimens from main haplotypes ensures that the cluster variation observed in a large proportion of specimens is covered. For example, Cluster 293 (Fig. 3) has 163 specimens. Checking only the extreme haplotypes would also reveal cluster failure, but the two checked specimens would represent the haplotypes of just 1.2% of the specimens (two singleton haplotypes). Including specimens from each of the two main haplotypes raises the percentage of specimens represented by the checked haplotypes to 88.3%. Therefore, our LIT protocol recommends studying specimens representing the main and the most distant haplotypes for PI clusters. Indeed, if feasible, it may be prudent to also check all main haplotypes for large non-PI clusters. Specimen selection can be aided by haplotype networks. These networks allow for a quick visual overview of the structure and patterns of variation within a cluster and facilitate the understanding of the molecular variation in the taxa.

### *Moving beyond Swedish Phorids*

Our study used data for ~18,000 specimens of scuttle flies (Diptera: Phoridae). Phoridae is a classic example of a dark taxon, that is, a specimen- and species-rich group where species discovery and identification are wanting. In terms of abundance, the family comprises ca. 10% of the total catch in Malaise trap samples from Sweden (Karlsson et al. 2020). In terms of diversity, it is the genus *Megaselia* that renders Phoridae a dark taxon—none of the other ca. 270 genera come close to the richness of this genus, and most are comparatively well explored. At present, *Megaselia* contains approximately 1,700 described species, but the worldwide fauna may be two orders of magnitude larger (Srivathsan et al. 2019). LIT will have to be tested and adapted to other taxa to optimize the number of specimens to examine. For example, we here used a 3% threshold for initial clustering, but this threshold is unlikely to be optimal for all taxa, data sets, or workflows. High thresholds yield fewer clusters for checking, but more clusters will fail the congruence test and require the inspection of additional specimens for determining correct species boundaries. On the other hand, the use of low thresholds for clustering results in the splitting of morphospecies (Fig. 7, Supplementary Table S2). Species split across multiple clusters are more difficult to resolve because one must check carefully whether each cluster only contains one species AND closely related clusters represent different morphospecies. This is laborious and time-consuming. We therefore recommend the use of clustering thresholds that are somewhat higher than what may eventually be revealed as optimal in terms of match ratios. This can be determined by examining known distances for a taxon or by studying a subset of specimens thoroughly. For the Swedish phorids, using OC at 3% was such a setting. Any algorithm that yields a mixture of split and lumped clusters is less suitable for the application of LIT. In this sense, we do not recommend PTP.

Other parameters used in LIT also need adjustment to specific taxa and data sets. For example, we here used a 1.5% maximum p-distance and determined the stability of clusters by comparing results obtained between 1% and 3% to identify PI clusters. However, the most appropriate values may be different for other studies. Fortunately, our study reveals that the threshold can be optimized based on a small number of randomly chosen clusters. We here had to study only 200 additional specimens to optimize the PI parameters, and we predict that fine-tuning of LIT for other taxa will require even fewer additional specimens. One way to lower the workload is to redefine "main haplotype" more conservatively based on specialist preference. For example, a 10% threshold for main haplotypes may require checking up to ten specimens per cluster (if a cluster were perfectly spread across ten haplotypes), while a 20% cutoff would require checking only five. If too many clusters are revealed to be PI, one can increase the efficiency of LIT by using more "lenient thresholds" which may then result in a moderate number of multispecies clusters being overlooked. For example, for our phorid data using the "large p-distance criterion" designated 24 clusters as PI, but in the end, this criterion only found one additional incongruent cluster and one species complex.

### *Future Uses of LIT*

To evaluate LIT further, it should be tested on many other taxa and a wide variety of sample densities. We must determine how LIT protocol can be modified with increased sampling, as the effectiveness of DNA barcoding for delimiting species is dependent on both the depth of sampling and the breadth of geographic scale (Bergsten et al. 2012; Huemer et al. 2014; Ahrens et al. 2016). Data sets that are shallowly sampled or of limited geographic scope can often appear decisive even in species-rich environments like the tropics (Hajibabaei et al. 2006; Smith et al. 2008), but with ever-expanding data sets the complexity of species delimitation (Sites and Marshall 2004; De Queiroz 2007; Wiens 2007) will increase. An even more rigorous test of LIT thus requires more sampling. Most species in our study were separated by 3% or more but as sampling increases, the mean genetic distances between species will decrease and eventually barcodes may no longer separate closely related species. Even within our data set, we had some species (such as in Cluster 101, Fig. 4) separated by just 0.6% (2 bp). Such small genetic differences between morphospecies suggest using only barcodes for species delimitation will lead to widespread taxonomic error, and such errors will only increase with sampling (Meier et al. 2021). This is not a problem unique to barcodes, it is a problem with relying on any single data source—morphology is also most likely to fail in cases of closely related species. Until we have better data sets with worldwide sampling for many groups, we will not be able to confidently understand the relationship between sampling and the proportion of clusters that remain stable.

In this study, we used mini-barcodes (313 bp), as they are easily obtained on short-read platforms such as Illumina and have been shown to perform comparably to full-length barcodes for the identification and delimitation of species (Yeo et al. 2020). Recent and rapid advancements to Oxford Nanopore MinION hardware, software, and bioinformatic pipelines are quickly making this technology an affordable (<0.10 USD per specimen) alternative for obtaining full-length barcodes at a large scale (Srivathsan et al. 2018, 2019, 2021). LIT can easily be implemented on data sets using full-length, or various mini-barcodes (but see Yeo et al. 2020 for a cautionary note on the use of some mini-barcodes).

The obvious next steps after species delimitation will be the description of new species and the identification of specimens that belong to described species. We predict that LIT will facilitate the description of taxa by yielding species units that have been delimited and validated with two data sources. Eventually, all data acquired for LIT can be automatically compiled into descriptions and diagnoses. The most time-consuming step for Holarctic taxa like Swedish species may very well be the process of determining which of the units already have a scientific name and which are new to science. Fortunately, advances in museomics are rapidly changing this situation, as we become able to reliably sequence old type material to match to new samples. In tropical regions, where the majority of species will be found and where the fauna of dark taxa is largely undescribed, most species could be described and named as new immediately after LIT has been completed (Dayrat 2005; Padial et al. 2010; Schlick-Steiner et al. 2010). As an example, in a previous project on Afrotropical phorids, we found evidence that a single site is home to >1,000 phorid species, but only 466 phorid species are described for the entire region (Srivathsan et al. 2019). In this case, describing and naming new species would face little delay.

## Conclusion

LIT is designed for dark taxa that contain large numbers of (mostly undescribed) species because traditional methods are not designed for taxa that are abundant and species rich. The method is not only important for applying integrative taxonomy to such taxa but also for the interpretation of metabarcoding data that are often derived from mass samples that are dominated by specimens belonging to dark taxa. We here develop and formalize a LIT approach using *COI* barcodes, which integrates the power and objectivity of large-scale barcoding with the kind of taxon-specific morphological expertise that only experts have, but that they cannot apply to many thousands of specimens. Future iterations of LIT could utilize other data sources, thus rendering a broad spectrum of dark taxa accessible for taxonomic work, including those outside Arthropoda (e.g., fungi, algae, etc.), for which barcodes will not be the ideal first data source. Dark taxa are likely critical to the functioning of our natural (and, ultimately, economic, and societal) systems; we must be committed to discovering and accurately identifying them.

## Supplementary Material

Supplementary material is available on GitHub at https://github.com/ronquistlab/taxon-cluster-paper.

## Acknowledgments

We would like to thank colleagues who provided valuable comments on the manuscript, including Brian Brown for his input on an early draft. For the Palaearctic phorid samples, we would like to thank Dave Karlsson, the Swedish Insect Inventory Project, and the crew at Station Linné. We would also like to thank Wan Ting Lee for help with molecular work, and the numerous staff, students and interns who have contributed to the establishment of the pipeline in the NUS laboratory.

## Funding

This work was supported by a Ministry of Education - Singapore grant on biodiversity discovery [R-154-000-A22-112] and a grant on the Megaselia of Sweden from Artdatabanken, Sveriges Lantbruksuniversitet [2016-203 4.3 to E.H.].

## References

Ahrens D., Fujisawa T., Krammer H.-J., Eberle J., Fabrizi S., Vogler A.P. 2016. Rarity and incomplete sampling in DNA-based species delimitation. Syst. Biol. 65:17.

Andersen A., Simcox D.J., Thomas J.A., Nash D.R. 2014. Assessing reintroduction schemes by comparing genetic diversity of reintroduced and source populations: a case study of the globally threatened large blue butterfly (Maculinea arion). Biol. Conserv. 175:34–41.

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. Syst. Biol. 61:851–869.

Bickel D. 2009. Why *Hilara* is not amusing: the problem of open-ended taxa and the limits of taxonomic knowledge. Diptera diversity: status, challenges, and tools. Leiden, Netherlands: E. J. Brill. p. 279–301.

Blaxter M.L. 2004. The promise of a DNA taxonomy. Philos. Trans. R. Soc. Lond. B. Biol. Sci. 359:669–679.

Butcher B.A., Smith M.A., Sharkey M.J., Quicke D.L.J. 2012. A turbo-taxonomic study of Thai Aleiodes (Aleiodes) and Aleiodes (Arcaleiodes) (Hymenoptera: Braconidae: Rogadinae) based largely on COI barcoded specimens, with rapid descriptions of 179 new species. Zootaxa 3457:1–232.

Cesari M., Giovannini I., Bertolani R., Rebecchi L. 2011. An example of problems associated with DNA barcoding in tardigrades: a novel method for obtaining voucher specimens. Zootaxa 3104:42.

Chapman A. 2009. Numbers of living species in Australia and the world. Canberra, Australia: Australian Biological Resources Study.

Curtis T. 2006. Microbial ecologists: it's time to "go large." Nat. Rev. Microbiol. 4:488–488.

Dayrat B. 2005. Towards integrative taxonomy. Biol. J. Linn. Soc. 85:407–415.

De Queiroz K. 2007. Species concepts and species delimitation. Syst. Biol. 56:879–886.

Disney R.H.L. 2009. Scuttle flies (Diptera: Phoridae) Part II: the genus *Megaselia*. Fauna Arab. 24:249–357.

Eberle J., Ahrens D., Mayer C., Niehuis O., Misof B. 2020. A plea for standardized nuclear markers in metazoan DNA taxonomy. Trends Ecol. Evol. 35:336–345.

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. Mol. Ecol. Resour. 13:851–861.

Hajibabaei M., Janzen D.H., Burns J.M., Hallwachs W., Hebert P.D.N. 2006. DNA barcodes distinguish species of tropical Lepidoptera. Proc. Natl. Acad. Sci. USA 103:968–971.

Hartop E.A., Brown B.V. 2014. The tip of the iceberg: a distinctive new spotted-wing Megaselia species (Diptera: Phoridae) from a tropical cloud forest survey and a new, streamlined method for Megaselia descriptions. Biodivers. Data J. 2:e4093.

Hartop E.A., Brown B.V., Disney R.H.L. 2016. Flies from L.A., The sequel: twelve further new species of Megaselia (Diptera: Phoridae) from the BioSCAN project in Los Angeles (California, USA). Biodivers. Data J. 4:e7756.

Hausmann A., Krogmann L., Peters R., Rduch V., Schmidt S. 2020. GBOL III: dark taxa. Available from: https://ibol.org/barcodebulletin/research/gbol-iii-dark-taxa/.

Hebert P.D., Cywinska A., Ball S.L., deWaard J.R. 2003. Biological identifications through DNA barcodes. Proc. Biol. Sci. 270:313–321.

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 19:219.

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:e115774.

Kapli P., Lutteropp S., Zhang J., Kobert K., Pavlidis P., Stamatakis A., Flouri T. 2017. Multi-rate Poisson tree processes for single-locus species delimitation under maximum likelihood and Markov Chain Monte Carlo. Bioinformatics 33:1630–1638.

Karlsson D., Hartop E.A., Forshage M., Jaschhof M., Ronquist F. 2020. The Swedish Malaise Trap Project: a 15 year retrospective on a countrywide insect inventory. Biodivers. Data J. 8:e47255.

Katoh K., Standley D.M. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol. Biol. Evol. 30:772–780.

Kekkonen M., Hebert P.D.N. 2014. DNA barcode-based delineation of putative species: efficient start for taxonomic workflows. Mol. Ecol. Resour. 14:706–715.

Kumar S., Stecher G., Li M., Knyaz C., Tamura K. 2018. MEGA X: molecular evolutionary genetics analysis across computing platforms. Mol. Biol. Evol. 35:1547–1549.

Kwong S., Srivathsan A., Vaidya G., Meier R. 2012. Is the COI barcoding gene involved in speciation through intergenomic conflict? Mol. Phylogenet. Evol. 62:1009–1012.

Larsen B.B., Miller E.C., Rhodes M.K., Wiens J.J. 2017. Inordinate fondness multiplied and redistributed: the number of species on earth and the new pie of life. Q. Rev. Biol. 92:229–265.

Leigh J.W., Bryant D. 2015. PopART: full-feature software for haplotype network construction. Methods Ecol. Evol. 6:1110–1116.

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. Front. Zool. 10:34.

Locey K.J., Lennon J.T. 2016. Scaling laws predict global microbial diversity. Proc. Natl. Acad. Sci. USA 113:5970–5975.

Losey J.E., Vaughan M. 2006. The economic value of ecological services provided by insects. BioScience 56:311.

Lücking R., Forno M.D., Moncada B., Coca L.F., Vargas-Mendoza L.Y., Aptroot A., Arias L.J., Besal B., Bungartz F., Cabrera-Amaya D.M., Cáceres M.E.S., Chaves J.L., Eliasaro S., Gutiérrez M.C., Hernández Marin J.E., de los Ángeles Herrera-Campos M., Holgado-Rojas M.E., Jonitz H., Kukwa M., Lucheta F., Madriñán S., Marcelli M.P., de Azevedo Martins S.M., Mercado-Díaz J.A., Molina J.A., Morales E.A., Nelson P.R., Nugra F., Ortega F., Paredes T., Patiño A.L., Peláez-Pulido R.N., Pérez R.E.P., Perlmutter G.B., Rivas-Plata E., Robayo J., Rodríguez C., Simijaca D.F., Soto-Medina E., Spielmann A.A., Suárez-Corredor A., Torres J.-M., Vargas C.A., Yánez-Ayabaca A., Weerakoon G., Wilk K., Pacheco M.C., Diazgranados M., Brokamp G., Borsch T., Gillevet P.M., Sikaroodi M., Lawrey J.D. 2016. Turbo-taxonomy to assemble a megadiverse lichen genus: seventy new species of Cora (Basidiomycota: Agaricales: Hygrophoraceae), honouring David Leslie Hawksworth's seventieth birthday. Fungal Divers. 84:139–207.

Meier R., Blaimer B.B., Buenaventura E., Hartop E., Rintelen T., Srivathsan A., Yeo D. 2021. A re-analysis of the data in Sharkey et al.'s (2021) minimalist revision reveals that BINs do not deserve names, but BOLD Systems needs a stronger commitment to open science. Cladistics 38:264–275.

Meier R., Shiyang K., Vaidya G., Ng P.K. 2006. DNA barcoding and taxonomy in Diptera: a tale of high intraspecific variability and low identification success. Syst. Biol. 55:715–28.

Meier R., Wong W., Srivathsan A., Foo M. 2016. $1 DNA barcodes for reconstructing complex phenomes and finding rare species in specimen-rich samples. Cladistics 32:100–110.

Meier R., Zhang G., Ali F. 2008. The use of mean instead of smallest interspecific distances exaggerates the size of the "barcoding gap" and leads to misidentification. Syst. Biol. 57:809–813.

Mora C., Tittensor D.P., Adl S., Simpson A.G., Worm B. 2011. How many species are there on Earth and in the ocean? PLoS Biol. 9:e1001127.

Padial J.M., Miralles A., De la Riva I., Vences M. 2010. The integrative future of taxonomy. Front. Zool. 7:16.

Page R. 2011. Dark taxa: GenBank in a post-taxonomic world. Available from: https://iphylo.blogspot.com/2011/04/dark-taxa-genbank-in-post-taxonomic.html.

Page R.D. 2016. DNA barcoding and taxonomy: dark taxa and dark texts. Philos. Trans. R Soc. Lond. B Biol. Sci. 371.

Pante E., Schoelinck C., Puillandre N. 2015. From integrative taxonomy to species description: one step beyond. Syst. Biol. 64:152–160.

Pentinsaari M., Salmela H., Mutanen M., Roslin T. 2016. Molecular evolution of a widely-adopted taxonomic marker (COI) across the animal tree of life. Sci. Rep. 6:35275.

Pérez-Silva J.G., Araujo-Voces M., Quesada V. 2018. nVenn: generalized, quasi-proportional Venn and Euler diagrams. Bioinformatics 34:2322–2324.

Puillandre N., Lambert A., Brouillet S., Achaz G. 2012. ABGD, automatic barcode gap discovery for primary species delimitation. Mol. Ecol. 21:1864–1877.

Puillandre N., Modica M.V., Zhang Y., Sirovich L., Boisselier M.-C., Cruaud C., Holford M., Samadi S. 2012. Large-scale species delimitation method for hyperdiverse groups: LARGE-SCALE SPECIES DELIMITATION. Mol. Ecol. 21:2671–2691.

Ratnasingham S., Hebert P.D. 2013. A DNA-based registry for all animal species: the barcode index number (BIN) system. PLoS One 8:e66213.

Riedel A., Sagata K., Suhardjono Y.R., Tänzler R., Balke M. 2013. Integrative taxonomy on the fast track - towards more sustainability in biodiversity research. Front. Zool. 10:1–9.

Riksförbundet Svensk Trädgård. 2018. Zonkartan. Available from: http://www.tradgard.org/svensk_tradgard/zonkartan.html.

Schlick-Steiner B.C., Steiner F.M., Seifert B., Stauffer C., Christian E., Crozier R.H. 2010. Integrative taxonomy: a multisource approach to exploring biodiversity. Annu. Rev. Entomol. 55:421–438.

Sites J.W., Marshall J.C. 2004. Operational criteria for delimiting species. Annu. Rev. Ecol. Evol. Syst. 35:199–227.

Smith M.A., Rodriguez J.J., Whitfield J.B., Deans A.R., Janzen D.H., Hallwachs W., Hebert P.D.N. 2008. Extreme diversity of tropical parasitoid wasps exposed by iterative integration of natural history, DNA barcoding, morphology, and collections. Proc. Natl. Acad. Sci. USA 105:12359–12364.

Sović I., Šikić M., Wilm, A. Fenlon, S.N., Chen, S., Nagarajan, N. (2016). Fast and sensitive mapping of nanopore sequencing reads with GraphMap. Nat. Commun. 7:11307.

Srivathsan A., Baloŏlu B., Wang W., Tan W.X., Bertrand D., Ng A.H.Q., Boey E.J.H., Koh J.J.Y., Nagarajan N., Meier R. 2018. A MinION$^{TM}$-based pipeline for fast and cost-effective DNA barcoding. Mol. Ecol. Resour. 18:1035–1049.

Srivathsan A., Hartop E.A., Puniamoorthy J., Lee W.T., Kutty S.N., Kurina O., Meier R. 2019. Rapid, large-scale species discovery in hyperdiverse taxa using 1D MinION sequencing. BMC Biol. 17:96.

Srivathsan A., Lee L., Katoh K., Hartop E., Kutty S.N., Wong J., Yeo D., Meier R. 2021. ONTbarcoder and MinION barcodes aid biodiversity discovery and identification by everyone, for everyone. BMC Biol. 19:217.

Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics.

Stork N.E. 2018. How many species of insects and other terrestrial arthropods are there on earth? Annu. Rev. Entomol. 63:31–45.

Schelske O., Retsa, A., Wilke B., Rutherford G., Jong R. 2020. Biodiversity and Ecosystem Services A business case for re/insurance. Available from: https://www.swissre.com/dam/jcr:a7fe3dca-c4d6-403b-961c-9fab1b2f0455/swiss-re-institute-expertise-publication-biodiversity-and-ecosystem-services.pdf

Tautz D., Arctander P., Minelli A., Thomas R.H., Vogler A.P. 2003. A plea for DNA taxonomy. Trends Ecol. Evol. 18:70–74.

Thomas J.A. 1995. The ecology and conservation of Maculinea arion and other European species of large blue butterfly. In: Pullin A.S., editor. Ecology and conservation of butterflies. Dordrecht: Springer Netherlands. p. 180–197.

Townes H. 1972. A light-weight Malaise trap. Entomol. News 83:239–247.

Truett G.E., Heeger P., Mynatt R.L., Truett A.A., Walker J.A., Warman M.L. 2000. Preparation of PCR-quality mouse genomic DNA with hot sodium hydroxide and tris (HotSHOT). BioTechniques 29:52–54.

Vaser R., Sović I., Nagarajan N., Šikić M. 2017. Fast and accurate de novo genome assembly from long uncorrected reads. Genome Res. 27:737–746.

Vitecek S., Kuèinić M., Previšić A., Živić I., Stojanović K., Keresztes L., Bálint M., Hoppeler F., Waringer J., Graf W., Pauls S.U. 2017. Integrative taxonomy by molecular species delimitation: multi-locus data corroborate a new species of Balkan Drusinae micro-endemics. BMC Evol. Biol. 17:129.

Vogler A.P., Monaghan M.T. 2007. Recent advances in DNA taxonomy. J. Zool. Syst. Evol. Res. 45:1–10.

Wang W.Y., Srivathsan A., Foo M., Yamane S., Meier R. 2018. Sorting specimen-rich invertebrate samples with cost-effective NGS barcodes: validating a reverse workflow for specimen processing. Mol. Ecol. Resour. 18:490–501.

Wiens J.J. 2007. Species delimitation: new approaches for discovering diversity. Syst. Biol. 56:875–878.

Will K.W., Mishler B.D., Wheeler Q.D. 2005. The perils of DNA barcoding and the need for integrative taxonomy. Syst. Biol. 54:844–851.

Wührl L., Pylatiuk C., Giersch M., Lapp F., von Rintelen T., Balke M., Schmidt S., Cerretti P., Meier R. 2021. DiversityScanner: robotic handling of small invertebrates with machine learning methods. Mol. Ecol. Resour. 22:1626–1638.

Yeo D., Srivathsan A., Meier R. 2020. Longer is not always better: optimizing barcode length for large-scale species discovery and identification. Syst. Biol. 69:999–1015.

Yong E. 2009 How research saved the Large Blue butterfly. Available from: https://www.nationalgeographic.com/science/article/how-research-saved-the-large-blue-butterfly.

Zhang J., Kapli P., Pavlidis P., Stamatakis A. 2013. A general species delimitation method with applications to phylogenetic placements. Bioinformatics 29:2869–2876.

Zhang J., Kobert K., Flouri T., Stamatakis A. 2014. PEAR: a fast and accurate illumina paired-end reAd mergeR. Bioinformatics 30:614–620.