“Dark taxonomy”: a new protocol for overcoming the taxonomic impediments for dark taxa and broadening the taxon base for biodiversity assessment

Rudolf Meier[1]
Amrita Srivathsan[1]
Sarah Siqueira Oliveira[2]
Maria Isabel P.A. Balbi[3]
Yuchen Ang[4]
Darren Yeo[5]
Dalton de Souza Amorim[3]

[1]Center for Integrative Biodiversity Discovery, Leibniz Institute for Evolution and Biodiversity Science, Museum für Naturkunde Berlin
[2]Departamento de Ecologia, Instituto de Ciências Biológicas, Universidade Federal de Goiás
[3]Departamento de Biologia, FFCLRP, Universidade de São Paulo
[4]Lee Kong Chian Natural History Museum, National University of Singapore
[5]Centre for Wildlife Forensics, National Parks Board Singapore

Corresponding author: Rudolf Meier (Rudolf.Meier@mfn.berlin)

Abstract.

We are entering the 6th mass extinction event on the planet with scarcely any data for "dark taxa" that comprise most animal species. These taxa have been neglected, because conventional taxonomic methods are not well-equipped to process tens of thousands of specimens belonging to thousands of species. We here test a new protocol for tackling this data deficiency problem by simultaneously addressing (1) taxonomic impediments, (2) lack of biodiversity baseline data, (3) and low impact of revisionary research. We propose to overcome the taxonomic impediments by carrying out revisions at small geographic scales to keep the number of specimens and species manageable. To lower the risk of taxonomic error, species delimitation and description are based on multiple data sources ("integrative taxonomy"). Secondly, we propose to create baseline data for dark taxa by initially revising them only based on fresh specimens obtained with the same sampling methods that are also used for biodiversity monitoring. This ensures that the revision constitutes a biodiversity baseline and the species most relevant for biomonitoring are revised first. Thirdly, we propose to improve the impact of taxonomic revisions by publishing two papers addressing different readerships (general and specialists). We illustrate our proposals by carrying out a taxonomic revision of the fungus gnats (Diptera: Mycetophilidae) of Singapore based on specimens obtained with Malaise traps placed at 107 sites across different habitats. We show that a first batch of specimens (N=1,456) contains 120 species, of which 115 are new to science and described in a separate taxonomic monograph. Species delimitation started with obtaining NGS barcodes (313-bp) that were used for estimating the number of MOTUs. This revealed 116-129 MOTUs depending on whether PTP, ABGD (P=0.001-0.060), or objective clustering (2-5%) was used. MOTU boundaries were then revised with "LIT" (Large-scale Integrative Taxonomy) which optimizes the integration of morphological and molecular data by using the molecular data. LIT revealed a match ratio of 91% for 3% MOTUs, but an even higher match ratio of 97% was found for MOTUs obtained with 5% Objective Clustering and ABGD (P=0.060). Using these parameters, only two MOTUs had to be fused for perfect congruence between molecular and morphological data for the 120 species. To test the completeness of our monograph, we then analyzed a second batch of 1,567 specimens. All clustering algorithms and settings revealed only 20 additional MOTUs; i.e., the first batch of specimens already covered 85% of the diversity, although 22% and 10% of the species are singletons or doubletons, respectively. Remarkably, >97% of all specimens in the second batch belonged to species described based on the first batch. The study thus demonstrates that revisions of dark taxa at moderate geographic scales with a moderate number of specimens from standardized traps can move dark taxa from being unavailable for biomonitoring to being well suited. In addition, the study increases the number of described species of Mycetophilidae in the entire Oriental Region by >25% indicating the size of taxonomic impediments for dark taxa in Southeast Asia.

Key words: Integrative Taxonomy, Dark Taxa, Reverse Taxonomic Workflow, Oriental Insect Fauna, NGS barcodes, Mycetophilidae

Introduction

Biodiversity assessment and monitoring are among the biggest and most urgent challenges of modern biology, given that many natural environments are disappearing fast and biodiversity loss is destabilizing whole ecosystems. However, due to taxonomic impediments, biodiversity assessment is difficult for many arthropod clades (e.g., Wheeler et al. 2004; Evenhuis 2007; de Carvalho et al. 2007). The problem is particularly severe in the tropics and for taxa characterized by high species diversity and abundance. For such taxa, conventional approaches to taxonomy involving morphospecies sorting tend to be ineffective, because they were not designed with hyperabundant and hyperdiverse taxa in mind (Hartop et al. 2022). Taxonomists have long been aware of these constraints and, thus, focused mostly on projects covering only moderate numbers of species and dealt with a small number of specimens pertaining to taxa that are overall already well-known (i.e., with prior revisions, identification keys, illustrations etc.). Projects that satisfied these criteria have been readily available and selecting them has been good career advice for young taxonomists. However, this also meant that many truly species-rich and abundant taxa remained neglected (Hartop et al. 2022; Srivathsan et al. 2023). Arguably, this has created a special kind of taxonomic impediment ("dark taxon impediment": Meier et al. 2022). Addressing this impediment has become high priority because taxonomic chauvinism (Bonnet et al. 2002; Troudet et al. 2017) needs to be overcome in an era where we need quantitative data on all of biodiversity.

What is needed are new taxonomic techniques that are scalable. Some examples illustrate how scalable they have to be: a recent study barcoded 7,059 specimens of scuttle flies (Diptera: Phoridae) collected by one Malaise trap over only eight weeks in a Ugandan National Park and found evidence for >650 species (Srivathsan et al. 2019). Yet, only 462 scuttle fly species have been described for the entire Afrotropical region (Phorid Catalog, 2023: accessed 13 August 2023). Similarly, the diversity of the gall midges (Diptera: Cecidomyiidae) has been estimated to be 800+ species at a single site of Costa Rica (Brown et al. 2018, Borkent et al. 2018) and a species diversity close to 8,500 was estimated for Canada based on barcodes (Hebert et al. 2016), although fewer than 250 Nearctic species have been described for this family (Savage et al. 2019). These cases illustrate why truly hyperdiverse taxa are sometimes called "open-ended." They will frustrate attempts at carrying out traditional comprehensive taxonomic revisions at reasonable geographic scales (Bickel 2009). An alternative term that is nowadays also frequently used for such clades is "dark taxa" (Hausmann et al. 2020, Hartop et al. 2022; Page 2011, 2016), which Hartop et al.

(2022) suggested should be restricted to clades where the undescribed fauna is estimated to exceed the described fauna by at least one order of magnitude and the total diversity exceeds 1000 species worldwide. These taxa need a different set of taxonomic protocols, which we here propose to call "dark taxonomy" to highlight that these techniques are fundamentally different from what is practiced for less diverse/abundant taxa.

The first key principle of "dark taxonomy" is embracing "faunistic sampling", i.e., the geographic scope of a taxonomic revision is determined by the number of specimens and species that can be reasonably handled and not by the putative range of species. Implicitly, almost all "turbo-taxonomy" studies have embraced such sampling without much discussion (e.g., Fernandez-Triana et al. 2014, 2022; Marsh et al., 2013; Riedel et al., 2013a,b; Riedel and Narakusumo 2019, but see Dijkstra et al. 2015). The lack of open discussion is presumably because it goes against the traditional recommendation that taxonomic revisions should cover all known species/specimens of a clade at a geographic scale appropriate for the species included in the clade. For example, if many species in a clade are found throughout the Palaearctic, one would carry out the revision at this geographic scale in order to properly assess species diversity and intraspecific variability. Such "completism" is desirable, but it contributes to taxonomic chauvinism because it cannot be applied to dark taxa. Two ways to relax the requirements is to limit taxonomic treatments to a small geographic area ("faunistics") and/or only work on a subset of the available specimens. Our revision of Mycetophilidae (Diptera; fungus gnats) is an extreme example for both. We only revise the fauna of a small island (Singapore, 730 km^2) and only used recently collected specimens obtained with Malaise traps.

Faunistic sampling carries obvious risks. The most important one is incorrect species delimitation because intraspecific genetic variability is poorly assessed for wide-spread species (Bergsten et al. 2012). However, we predict that this error is manageable by applying the principles of integrative taxonomy; i.e., species delimitation based on multiple data sources (Dayrat 2005). It is likely that containment will be particularly effective if the characters used for species delimitation were generated by different evolutionary processes. With regard to morphology, taxonomists usually collect information from a broad array of structures including genitalia, legs, wings etc. Genitalia are likely to be under strong sexual selection (e.g., Eberhard 1985, Rowe and Arnqvist 2012), while many other body parts are presumably more shaped by viability selection. It thus appears likely that most speciation events will leave traces in such a diverse array of structures, regardless of whether speciation was fast or slow. However, there are exceptions, as is well known based on the literature on "cryptic species" (Bickford et al. 2007). Therefore, morphology should be paired with a type of data that evolves largely via drift and thus reflects time of divergence. This is the case for DNA barcodes, given that closely related species mostly differ with regard to

synonymous sites (Kwong et al. 2012; Pentinsaari et al. 2016). Large DNA barcode distances can therefore flag cases where morphological differentiation is lower than expected based on time of divergence. This can then trigger more detailed work to resolve species limits for taxa that have high intra- or low interspecific variability ("reciprocal illumination"). We have repeatedly found this to be very effective at resolving taxonomic problems for a range of fly families (Ang et al. 2013, Tan et al. 2010, Rohner et al. 2014, Laamanen et al. 2003).

The second element for tackling large numbers of specimens is what we call the "reverse workflow". Traditionally, species discovery has started with sorting specimens to species using morphology. Subsequently, the morphospecies were validated with few barcodes. This workflow is ineffective for dark taxa, where even samples collected at a small geographic scales routinely contain too many specimens and species for morphospecies sorting. Imagine having to sort the aforementioned >7000 specimens of Ugandan phorids into 650 species in this manner, or the 18,000 phorids in Hartop et al. (2022) into >350 species. We have thus argued before that such samples should first be sorted to Molecular Operational Taxonomic Units (MOTUs) with barcodes (Wang et al. 2018; Hartop et al. 2022). This is now feasible because barcodes can now be acquired semi-automatically and at low cost by staff who only specialize in molecular methods (Meier et al. 2016; Srivathsan et al. 2019, 2021). Presorting large numbers of specimens to putative species means that taxonomic experts can concentrate on revising MOTU boundaries using other sources of data (Wang et al. 2018; Hartop et al. 2022).

Delimiting large numbers of species based on two types of data also requires explicit rules for assessing and resolving conflict between data types. We recently proposed "Large-scale Integrative Taxonomy" (LIT) for this purpose (Hartop et al. 2022). It addresses the "integrative taxonomy conundrum" that collecting two different types of data for all specimens will slow down taxonomy when faster speed is needed to address taxonomic impediments. Fortunately, the conundrum can be overcome by speeding up collection of one type of data (e.g., semiautomatic acquisition of barcodes) and then only testing MOTUs delimited with these data by subsampling specimens for a second type of data (Hartop et al. 2022). This is what LIT implements. In the first stage, barcodes are used to generate a set of MOTUs based on one criterion (e.g., 3% clustering). In the second step they are classified as "stable" or "instable" depending on whether different algorithms and clustering thresholds change MOTU composition. In the third stage, the MOTUs are tested by studying the morphology of specimens representing the main and/or most divergent haplotypes—instable MOTUs are tested more extensively than stable ones (see Hartop et al. 2022). In the fourth stage, those MOTUs that that are congruent with morphological evidence are accepted as species. The fifth stage is reserved for resolving conflict between morphological and molecular evidence. As argued in Hartop et al. (2022), such conflict is expected for recently diverged species which are

expected to be incorrectly lumped into one MOTU (e.g., also due to introgression) and “old” species with diverged allopatric populations which will be divided into multiple MOTUs. Conflict can thus usually be resolved by testing whether MOTU boundaries obtained with different algorithms or clustering settings are congruent with morphological evidence. If so, these morphospecies can also be described because the molecular and morphological data are not genuinely in conflict. If not, the MOTUs/morphospecies are left unresolved until a third data source can be used to determine species boundaries. Hartop et al. (2022) showed that only 5-10% of all specimens had to be checked to find congruent species hypotheses for the ca. 350 MOTUs that were analyzed by studying 18,000 specimens.

Sorting specimens into putative species and revising the MOTU boundaries using LIT yields species-level units. The next step is establishing which species are already described and which need description. As illustrated in our companion monograph of mycetophilids of Singapore (Amorim et al., 2023), the task of sorting through old descriptions is manageable for dark taxa in the tropics because so few species have been described. The “superficial description impediment” of Meier et al. (2021) may thus be somewhat less onerous for the taxa and region of the planet that are most in need for dark taxonomy. In other parts of the world, revisions are more likely to get stuck after completing LIT because too many scientific names cannot be resolved. This is frustrating, but it is already important progress when all species have been delimited two types of data. Museomics is evolving quickly and accelerate the matching of delimited species without names with barcodes from reliably identified or type material (e.g., Santos et al. 2023).

Developing high-throughput taxonomic techniques for dark taxa is important, but we will also need more taxonomists to take on dark taxa. In proposing “dark taxonomy”, we suggest two ways for increasing impact. The first is that initial revisions of dark taxa at small geographic scales should be based on specimens that were collected with the kind of sampling techniques that are also used for biomonitoring biodiversity. For many groups of insects, this would be Malaise-, pitfall-, light-, and/or flight intercept traps. Revising taxa based on such samples will simultaneously yield biodiversity baselines and species descriptions that are immediately relevant for biomonitoring. Furthermore, the DNA barcode database generated becomes a valuable resource for metabarcoding and metagenomics projects. Fresh specimen samples are also easier to process at scale, because the specimens are suitable for rapid barcoding and locality information is usually available in a digital format. The second strategy is to publish the results of such revisions in two manuscripts. The first presents the results of interest to a broad readership, while the species descriptions can be presented in a second monograph targeting taxonomic experts. To ensure that the information in both manuscripts are available at the same time, we recommend joint publication as a preprint. It should be duly noted, however, that this requires the registration of all scientific names in Zoobank and the

inclusion of a disclaimer according to section 8.2. of the ICZN code that the preprint is not issued for the purpose of zoological nomenclature and should be considered not published within the meaning of the code.

In this paper, we propose and test "dark taxonomy" by first revising the fauna of Mycetophilidae of Singapore based on an initial batch of 1,456 specimens. We then test the efficiency of the protocol by using a second batch of 1,567 specimens to measure the proportion of species already described based on the first batch. In addition, we illustrate the size of the taxonomic impediment for dark taxa in the Oriental region by demonstrating that revising the fauna of a small island increases the number of described species by >25%.

Material and Methods

The project proceeded through five phases, but the first would not have to be repeated for many other dark taxa.

Phase 1: Collecting. Malaise traps were deployed for varying periods (2-6 months) between April 2012 and June 2019 at 107 sites in Singapore (Figs. 1A–B). The sites represent the following habitat types in Singapore: degraded urban secondary forest, old and maturing secondary forest, primary forest, coastal forest, swamp forest, freshwater swamp (lacking mature trees), and old-growth and replanted mangrove forests (Fig. 2A–E). The samples were collected weekly and preserved in 70% ethanol. Subsequently, they were sorted to order/family by parataxonomists and entomologists. Mycetophilids were present in most sites, though not particularly abundant in some

habitats (e.g., mangroves). Details on the sampling campaign are in Yeo et al. (2021).

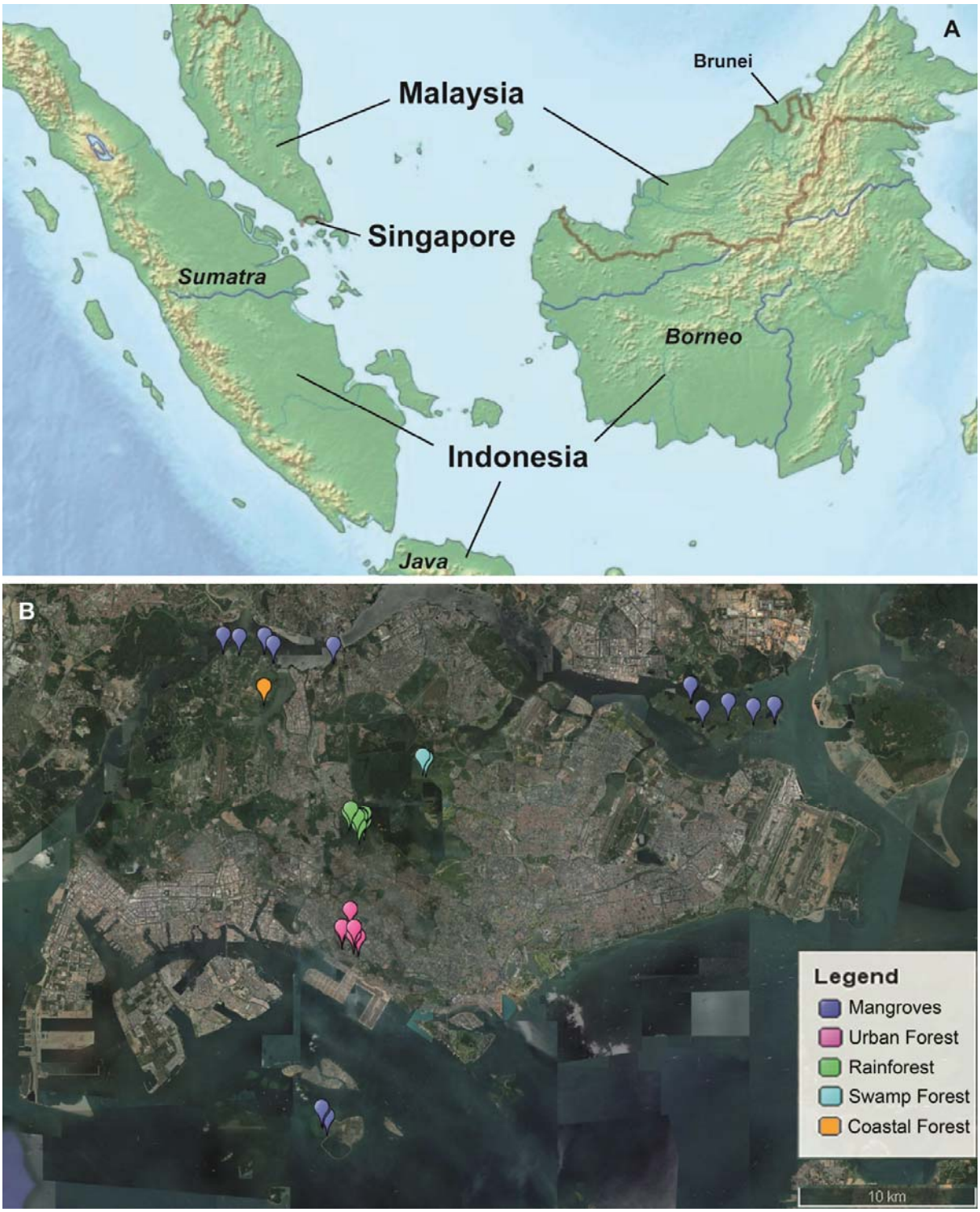

Figure 1. Singapore sites sampled with Malaise traps (modified from www.freeworldmaps.net). A. Southeast Asia, with relative position of Singapore. B. Collecting sites in different environments.

Figure 2. Environments sampled in Singapore. A: Mangrove, B: Rainforest, C: Urban forest, D: Swamp forest, E: Freshwater swamp.

Phase 2: Estimating the number of MOTUs. We processed two batches of specimens. The first had 1,456 and the second, 1,567 specimens. The latter was sequenced after the cut-off date for the descriptive monograph (Amorim et al. 2023). All specimens were sequenced for a 313-bp fragment

of the *cytochrome oxidase I* gene (*COI*) given that such minibarcodes amplify better than the full-length barcode (Srivathsan et al. 2021) and nevertheless yield similar species-level signal (Yeo et al. 2020). The minibarcode was amplified for each specimen using a protocol in Meier et al. (2016). In the early phases of the study, DirectPCR (Wong et al. 2014) was employed; i.e., 1 or 2 legs of each specimen were used as the DNA template for amplification. For specimens collected later, DNA extraction was performed by immersing the whole specimen in Lucigen's QuickExtract solution or HotSHOT buffer (Montero Pau et al. 2008). PCR was performed with the following primer pair mlCO1intF: 5'-GGWACWGGWTGAACWGTWTAYCCYCC-3' (Leray et al. 2013) and jgHCO2198: 5'-TANACYTCNGGRTGNCCRAARAAYCA-3' (Geller et al. 2013). These primers were labelled at the 5' end with 9-bp tags that differed from each other by at least three base pairs. Every specimen in each sequencing library was assigned a unique combination of forward and reverse primer tags. This allows for reads to be assigned to their specimen of origin. A negative control was used for each 96-well PCR plate to detect contamination. Amplification success was estimated via gel electrophoresis of a subsample of eight wells from each plate.

Amplicons were pooled at equal volume within each plate. Equimolar pooling across plates was performed by approximating concentration based on the presence and intensity of the gel bands. The pooled samples were cleaned with Bioline SureClean Plus before being sent for library preparation at AITbiotech using TruSeq Nano DNA Library Preparation Kits (Illumina) or the Genome Institute of Singapore (GIS) using NEBNext DNA Library Preparation Kits (NEB). Paired-end sequencing was performed on Illumina Miseq (2x300-bp or 2x250-bp) or Hiseq 2500 platforms (2x250-bp). The sequences were obtained across multiple runs, which allowed for troubleshooting and re-sequencing specimens which failed to yield a sufficiently high number of reads during the first sequencing attempt. Some of the specimens were also sequenced with MinION (Oxford Nanopore) using primers labelled with slightly longer tags (13-bp; see Srivathsan et al. 2019). The raw Illumina reads were processed with the bioinformatics and quality-control pipeline described in Meier et al. (2016). A BLAST search to NCBI GenBank's nucleotide (nt) database was also conducted to identify contaminant sequences by parsing the BLAST output with *readsidentifier* (Srivathsan et al. 2015). Barcodes with that matched non-target taxa at >97% identity were discarded.

Barcodes lacking stop codons were aligned with MAFFT v7 (Katoh and Standley 2013) using default parameters and analysed using three different species delimitation algorithms: objective clustering (Meier et al. 2006), Automatic Barcoding Gap Discovery (ABGD: Puillandre et al. 2012) and Poisson Tree Process (PTP: Zhang et al. 2013). Objective clustering used a python script that implements the objective clustering algorithm in Meier et al. (2006) and we obtained the results for four p-distance thresholds (2–5%) encompassing thresholds commonly used in the literature for

species delimitation (Ratnasingham and Hebert 2013). ABGD was performed using uncorrected p-distances and the minimum slope parameter (-X) 0.1, with the default range of priors (P = 0.001 – 0.1). Lastly, a maximum likelihood phylogeny was generated from the barcode sequences in RAxML v.8 (Stamatakis 2014) with the GTRGAMMA model and with 20 independent tree searches. The barcode tree was then used for the PTP analysis using the implementation provided in mPTP (--ml --single) (Kapli et al. 2017; Zhang et al. 2013) algorithms. As described in Wang et al. (2019), the vials with the specimens were then physically sorted into bags that corresponded to 3% p-distance clusters. Specimens from these bags were used for morphological verification of species boundaries according to the rules of LIT.

Phase 3: Integrative Taxonomy

We started the morphological study by slide mounting at least one male and one female (if available) for each 3% MOTU. The specimens were cleared with KOH, dehydrated in ethanol, dissected to separate wings, abdomens with terminalia, and head/thorax. The body parts were mounted in Euparal (modified from Walker and Crosby 1988; Huber and Reis 2011) under three separate coverslips. Particularly, the specimens that underwent DNA extraction with Proteinase K overall worked well because the proteins were already digested, while the exoskeleton was well preserved for slide mounting (Santos et al. 2018). However, some soft specimens collapsed, affecting the quality of slide mounts.

This initially screen was then enhanced to test the LIT rules outlined in Hartop et al. (2022). All non-singleton barcode clusters obtained at 3% were classified as stable or instable based on a stability index that quantifies whether cluster content changes when the clustering threshold is increased from 1% to 3%. The clusters where the maximum pairwise distance between any two specimens was >1.5% were also flagged as instable. All stable clusters were then tested by studying the morphology of a pair of specimens representing the most distant haplotypes, while for all instable clusters the morphology of one specimen representing the main haplotypes was also studied. For most genera, the diagnoses include male terminalia features, such that only clusters with males could provide sufficient information on conspecificity. For each cluster we then assessed whether (1) the examined specimens from the same MOTU belonged to the same morphospecies and (2) differed from the morphospecies in the “neighboring” clusters. If (1) was violated, the cluster was split to test whether morphospecies and DNA barcode information were congruent at a lower clustering threshold. If (2) was violated, it was tested whether fusing “neighboring” clusters restored congruence with morphospecies. Eventually, we accepted those grouping decisions based on DNA barcodes that were congruent with morphological information. No species delimitation

was performed when morphospecies were incongruent with all grouping statements based on barcodes.

Phase 4: Species identification and description

To decide whether a specimen belonged to a described species or one that had to be newly described, we extensively reviewed the literature on the Oriental fauna of Mycetophilidae, which is here also cited because it forms the foundation of all taxonomic work on the group (Meier, R. 2017). For the Oriental species described by Edwards (1925, 1926, 1927, 1928, 1929, 1931, 1933, 1935, 1940), we studied and photographed primary types deposited at the Natural History Museum, London. For more recently described Oriental species of Mycetophilidae, we were able to resolve species identity based on published careful descriptions with good illustrations (Colless 1966; Zaitzev 1982; Sivec and Plassmann 1982; Wu and Yang 1986; Bechev 1995; Søli 1996, 2002; Kallweit 1998; Matile 1999; Xu and Wu 2002; Wu et al. 2003; Papp 2004; Hippa et al. 2005; Hippa 2006, 2007, 2008, 2009, 2011; Ševčík 2001; Ševčík and Hippa 2010; Hippa and Ševčík 2010, 2013; Ševčík et al. 2011; Borkent and Wheeler 2012; Kurina and Hippa 2015; Hippa and Saigusa 2016; Hippa and Kurina 2018; Magnussen et al. 2019; Fitzgerald 2017; Kaspřák et al. 2017; Kjaerandsen et al. 2023). Some Oriental species, especially those from Sri Lanka, cannot be resolved because the descriptions are insufficiently detailed and the types were lost or are unavailable. We here assumed that the species from Singapore are allospecific based on geography, which is consistent with results found in South America, with well-delimited endemism for Mycetophilidae species (Amorim and Santos 2018). Please note that this preprint is not issued for the purpose of zoological nomenclature and should be considered not published within the meaning of the code (ICZN code, Section 8.2).

Phase 5: Completeness of the faunal assessment

The barcode data for the second batch of specimens (N=1,567) was used to test how complete the revision based on the first batch of specimens was (N=1,456). For this purpose, the genetic data were analyzed as described earlier for the first batch of sequences. Subsequently, we assessed how complete the initial revision was in the light of the species diversity represented in the second set of specimens. We also determined the number of new haplotypes, clusters, and how many specimens from batch two belonged to species described based on batch 1. We then compared these observed values to those obtained by randomized assignment of specimens to batch 1 and batch 2, assuming the same sample sizes (batch 1, n=1,456 and batch 2, n=1,567) in order to account for any effect of differences in sampling. This was done using a custom script

(https://github.com/asrivathsan/darktaxa). Lastly, we estimated how many specimens would have had to be studied if LIT was performed on both batches of specimens (N=3,023).

Results

Phase 1: Collecting

The material for this study was obtained during 3,526 Malaise trapping weeks distributed across five different habitats in Singapore: mangroves (74 sites, 2,162 trap weeks), tropical forest (nine sites, 567 weeks), urban forests (15 sites, 280 weeks), swamp forest (four sites, 262 weeks), coastal forest (10 sites, 156 weeks), and freshwater swamp (seven sites, 99 weeks). A total of 69 traps collected 3,030 mycetophilid specimens that were successfully sequenced. The overall number of mycetophilid specimens collected is somewhat lower than expected for such a large number of Malaise trap samples. This is due to fact that almost two thirds of the samples (61%) were obtained from mangroves that are comparatively poor for Mycetophilidae.

Phase 2: Molecular processing

The 1,456 COI barcodes from the first batch clustered into 116–129 MOTUs via objective clustering at 2 – 5% uncorrected p-distances (Table 1). With ABGD, the number of MOTUs range from 116–127. PTP clustered the barcodes into 129 MOTUs respectively. The 1,567 COI barcodes from the second batch yielded 99 MOTUs from 5% p-distance clustering.

Phase 3: Integrative Taxonomy

Congruence between morphospecies and MOTUs was assessed with match ratios as described in Ahrens et al. (2016). It was overall high (0.87-0.97), with the highest match ratio observed for 5% clusters and MOTUs obtained with ABGD at P=0.06 (0.97, 114 morphospecies congruent).

Table 1. Integrative species delimitation using different algorithms for MOTU estimation. The number of morphospecies with molecular data was 118.

Algorithm	PTP	Objective Clustering				ABGD				
		2%	3%	4%	5%	P=0.001 -0.005	P=0.008 -0.036	P=0.013 -0.022	P=0.036	P=0.06 0
No. MOTUs	129	129	125	122	116	127	126	125	122	116
No. congruent morphospecies	108	108	111	114	114	110	111	111	114	114
Split morphospecies	10	10	7	4	0	8	7	7	4	0
Fused morphospecies	0	0	0	0	2	0	0	0	0	2
Match ratio	87%	87%	91%	95%	97%	90%	91%	91%	95%	97%

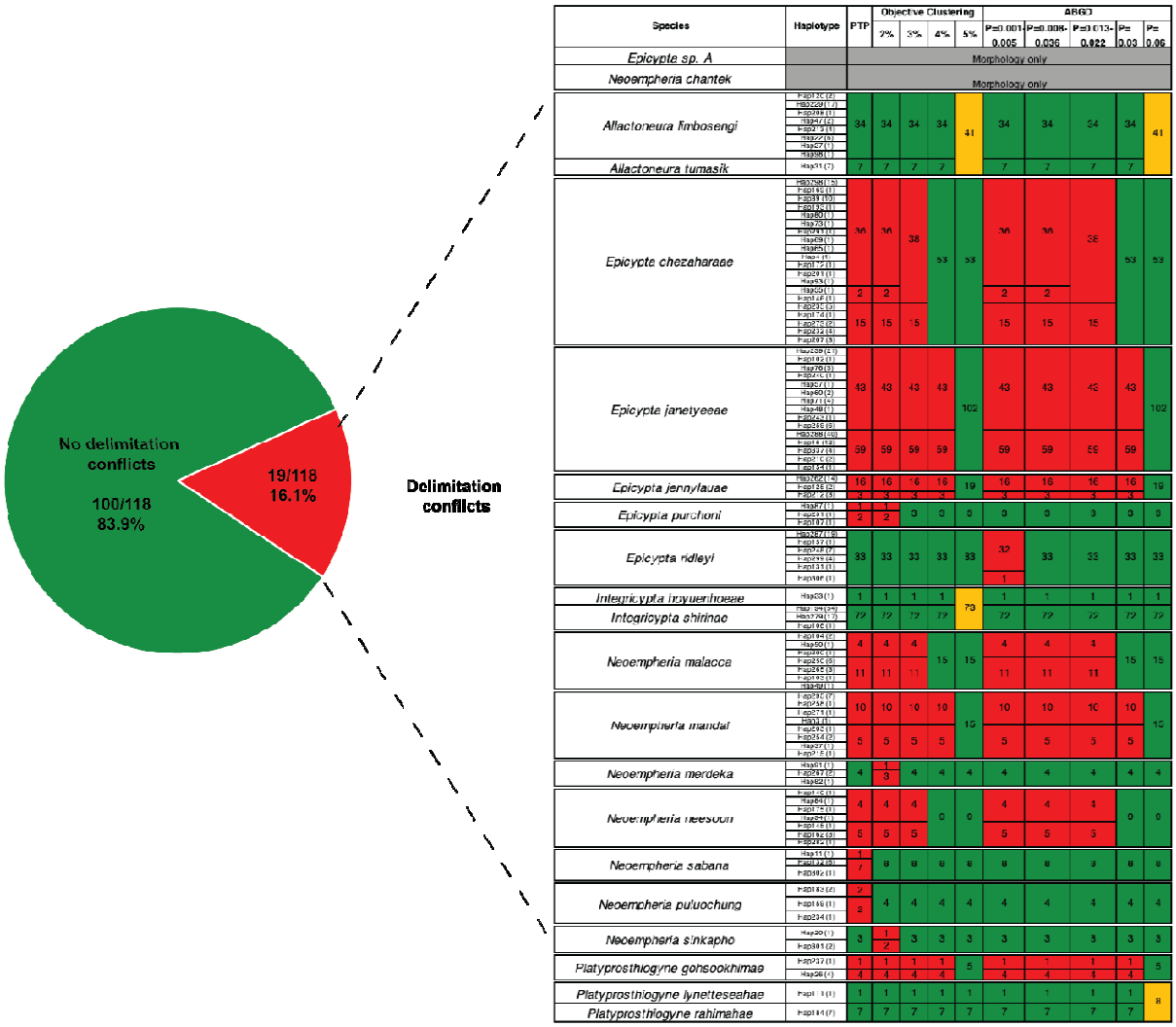

Figure 3. Most MOTUs are stable and congruent with morphospecies (green; N=100). Remaining species have splitting (red) and lumping (yellow) conflict with molecular data depending on species delimitation algorithm.

Phase 4: Species identification, description and naming

The application of LIT to the 125 MOTUs (3%) required morphological study of 225 specimens. Of the 125 MOTUs, 28 were singletons, 73 MOTUs were stable, and 24 were instable (530 specimens). One challenge with carrying out the LIT analysis is that for approx. 40% of haplotypes, only female specimens were available. This weakens the test for congruence with MOTU boundaries. Overall, congruence with morphospecies required fusing closely related MOTUs in seven cases (Figure 3; 7x2 MOTUs). Subsequently, 118 species were delimited and described belonging to 21 known and one new genus (note that the monograph also describes two additional species based on morphology only: Amorim et al. 2023). Only four species were already described

prior to this revision: *Eumanota racola* Søli, *Metanepsia malaysiana* Kallweit, *Parempheriella defectiva* Edwards, *Neoempheria dizonalis* Edwards. In addition, one species of *Chalastonepsia* Søli is only represented by females in our sample and may belong to one of the described species of the genus given that the species in our material has a 99% barcode match to a specimen identified as cf. *Chalastonepsia hokkaidensis*) in BOLD (BOLD:ACP9308). Note, however, that this identification in BOLD is problematic given that the record is of a single female from Malaysia, while *C. hokkaidensis* Kallweit was described from Japan. The remaining 113 new species with molecular and morphological data are described and illustrated in the companion taxonomic monograph (Amorim et al. 2023). Images are also available from "Biodiversity of Singapore" (https://singapore.biodiversity.online/taxon/A-Arth-Hexa-Dipt-Mycetophilidae).

Phase 5: Completeness of the faunal assessment

The second batch of specimens yielded belonged to 98 MOTUs at 5% in a combined analysis of all the 3,023 barcodes. 5% MOTUs were here used for clustering because it maximized congruence between morphology and molecular data for batch 1. Of these, 78 belonged to MOTUs already examined for species descriptions based on the specimens in the first batch (78/98 or 79.5% of species in batch 2). The newly discovered 20 MOTUs were all stable (no compositional changes with all delimitation algorithms and settings used here) and overall so rare that they only represented 36 of the 1,567 specimens in the second batch. Overall, more than 90% (97.7%) of all specimens in batch 2 belonged to MOTUs that were also found in batch 1 (Figure 4). In order to assess if the observed coverage of species and specimens for batch 2 was influenced by sampling, we compared these values to those obtained by randomized assignments of specimens to batch 1 and batch 2 (100 randomizations). Randomized assignments of specimens to batch 2 revealed coverage that closely reflected observed values: 81.6 ± 2.2% of MOTUs and 98.1 ± 0.4% of specimens would belong to MOTUs found in batch 1 (Figure 4). Thus, if batch 1 is well studied in terms of integrative taxonomy, very few specimens in batch 2 would belong to new 5% MOTUs and require additional taxonomic work.

At 5% objective clustering, the proportion of singletons was 22% (N=30) for all specimens, 22% (N=26) for dataset 1, and 25% (N=25) for dataset 2. The corresponding numbers for doubletons were 10% (N=14), 9% (N=11), and 14% (N=14). Carrying out a LIT analysis for all 3,023 specimens would have required checking 146 3% MOTUs of which 31 were singletons, 81 stable, and 34 were instable (1,781 specimens). Despite the large number of specimens in the instable clusters, only 286 specimens of the combined dataset would have to be checked to sample haplotypes according to LIT.

Table 2. MOTU (3%) and specimen abundance for the second batch of specimens and all data.

Dataset	**Batch 1**	**Batches 1 + 2**
3% MOTUs (No. of specimens)	125 (1456)	146 (3023)
Singleton MOTUs	28	31
Stable MOTUs (No. of specimens)	73 (898)	81 (1211)
Instable MOTUs (No. of specimens)	24 (530)	34 (1781)
LIT: No. of specimens to be examined (singleton/stable/instable)	225 (28/130/67)	280 (31/139/110)

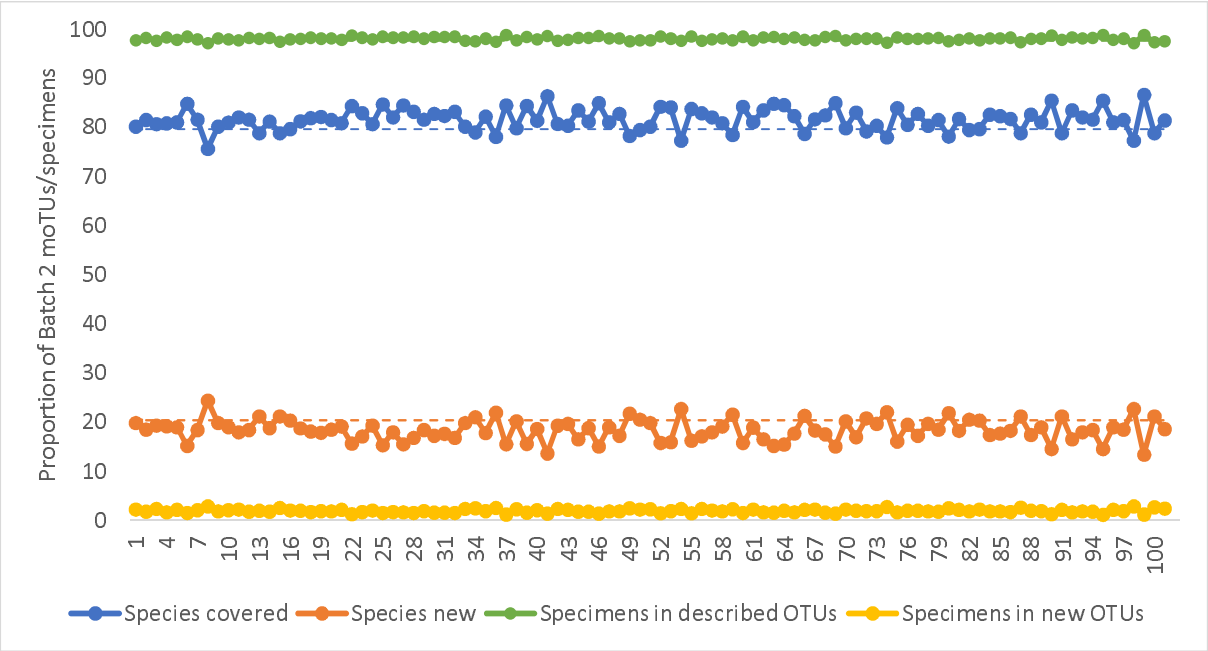

Figure 4. An overwhelming number of specimens and most species in specimen batch 2 belong to species described based on specimens in batch 1 (~80% of species and >97% of specimens). Dashed lines represent the observed values while the solid lines represent the results of 100 randomizations.

Discussion

Biodiversity loss is now considered one of the top five threats to planetary health (World Economic Forum 2023). Yet, most animal species are undescribed, unidentifiable, and lack baseline data for biomonitoring. Much of this diversity apparently belongs to dark taxa such as the top 20 species-rich families of flying insects that were identified in Srivathsan et al. (2023) based on samples from five biogeographical regions. For many of these taxa, we know how to obtain standardized quantitative samples (e.g., Malaise, pitfall, flight intercept traps), but most samples are

currently only evaluated with metabarcoding (Yu et al., 2012). Such approaches are unable to overcome taxonomic impediments because they disrupt the association between specimen and genetic data. In addition, metabarcoding only provides approximate abundance/biomass data. To overcome these problems, we propose that some samples for each site should be analyzed with "dark taxonomy" to properly delimit species, obtain species-specific barcode libraries, facilitate proper taxonomic documentation through description and naming, and generate biodiversity baseline data for biomonitoring.

We here show that dark taxonomy can fairly quickly yield a good estimate of species diversity and provide species names and barcodes for most common species belonging to a dark taxon. Our initial set of specimens consisted only of 1,456 specimens. They are here shown to belong to 118 species delimitated with morphological and molecular data. We then tested the completeness of the revision by using the data for an additional 1,567 specimens. Fortunately, the observed number of MOTUs increased by only 17% (N=20); i.e., the first richness estimate was more than 80% complete. Furthermore, 78 of the 118 species in the first batch were also represented in the second batch. Not surprisingly, this involved mostly common species, so that the proportion of specimens from the second batch that belong to species described in the first batch is very high (>97%). This means that one moderately scaled revision of a dark taxon was sufficient to ready a country like Singapore for the inclusion of this taxon (Mycetophilidae) in biomonitoring. This is impressive given that the number of known and described species of Mycetophilidae in Singapore now exceeds the number of species in major vertebrate groups (mammals: 83 spp.; amphibians: 26 spp.; "reptiles": 109 spp; source: Singapore National Parks Board [Red Data Book] Species List, 2023). Moreover, all the new species can be identified based either on barcodes or morphology and sampled minimally invasively with Malaise traps. Currently, very little is known about the natural history of tropical fungus gnats so it is difficult to assess whether they could be used as indicator taxa for habitat quality or fungus richness. However, this seems likely, given the dependency of Mycetophilidae on fungi and studies such as the ones by Økland (1994, 1996) that demonstrated that mycetophilids are good indicators of temperate forest quality. However, more environmental correlates need to be collected for the tropics and it will be important to associate larvae and adults (see Yeo et al. 2018) to understand to what extent mycetophilid species are specialized on specific species of fungi and how mycetophilid species-richness and abundance could be used as an indirect evidence of tropical forest health.

To fully appreciate the potential use of dark taxonomy, it also is important to consider that Srivathsan et al. (2023) showed that more than half of the specimens and species in Malaise trap samples belong to approximately 20 family-level dark taxa (see also Brown 2005, Karlsson et al. 2020). This means that applying a technique like "dark taxonomy" to a few complete Malaise trap

samples could simultaneously yield biodiversity baseline data for many numerically dominant families and hundreds of species simultaneously. This is realistic, because barcoding is now so easy and cheap (Srivathsan et al. 2021) that dark taxonomy is also suitable for biodiverse countries with lesser science funding. PCR costs can be as low as 0.05 USD per specimen, and sequencing with Illumina's NovaSeq costs only 1-2 cents/barcode (Srivathsan et al. 2021). A self-contained barcoding lab covering all steps from specimens to barcodes can be installed for <USD 10,000 if it utilizes Oxford Nanopore Technologies (ONT) sequencers (e.g., MinION; Srivathsan et al. 2021). The use of 3rd generation sequencing technologies raises the sequencing cost for DNA barcodes, but it still means that the consumable cost for barcoding the ~3,000 specimens in our study costs USD 150 (Illumina) to USD 375 (ONT). Specimen handling costs are also dropping with the use of robots for sample sorting (see Wührl et al. 2022). Carrying out taxonomic revisions for dark taxa will inevitably remain challenging but "dark taxonomy" can help with simultaneously addressing taxonomic impediments and the need for biodiversity baseline data.

Objections to dark taxonomy

The proposal of "dark taxonomy" is likely to be controversial for a variety of reasons. One is that specimens/types from dry collections will initially only be used if they are important for resolving species names, i.e., only barcodes for types or reliably identified specimens are important. However, many rare and most extinct species are only found in dry collections. They are rarely or never found in the kind of traps used for biomonitoring and thus not covered by "dark taxonomy" revisions. However, we would argue that the advantages of dark taxonomy outweigh these disadvantages. Concentrating on fresh material from standardized traps means that specimens are suitable for barcoding. Secondly, standardized samples come with complete metadata and there is no need to digitize thousands of specimen labels individually. None of this can be said for taxonomic revisions that deal with a mixture of wet and dry specimens. This does not mean that dry material will become irrelevant. Dark taxonomy is designed to rapidly improve taxonomic knowledge for a poorly known taxon to reach the level where conventional taxonomic approaches can be applied. At this stage, pinned museum specimens will become critical for testing uncertain species boundaries and estimating the full species-level diversity of a clade.

We also suspect that some systematists may be concerned about working at small geographic scales. However, there are two reasons why we are less worried. Firstly, we would like to point to the success of those turbo-taxonomic studies that seem to have found few cases, if any, where species boundaries had to be revisited later when the fauna of neighboring areas was covered. The best example is the weevil genus *Trigonopterus*, where several hundred species have been described in regional treatments (Narakusumo et al. 2019, 2020; Riedel 2022; Riedel and Narakusumo 2019;

Riedel and Tänzler 2016; Riedel et al. 2013b, 2014; Van Dam et al. 2016). Of course, only time will tell as to which proportion of species proposed based on poor geographic sampling will have to be revised as geographic sampling increases, but it appears unlikely that the proportion will be high. This prediction is based on the current observation that barcodes alone tend to successfully delimit the majority of species. Note, however, that the precise proportion remains unknown given that most species still lack barcodes and comparatively few species have been sampled across their full geographic range. Denser sampling tends to obscure the signal in barcode data as can also be seen in our study given that the proportion of specimens in instable MOTUs increased from 530 in dataset 1 (36%) to 1,781 in the combined dataset (59%). However, overall, we are optimistic that the vast majority of species delimited with dark taxonomy will be stable, because the species boundaries are determined by morphological and molecular data. So, both data would have to be misleading to generate incorrect species limits.

Those systematists who are still reluctant to accept that dark taxa should be revised at small geographic scales may want to consider the alternative. By describing more than 100 species based on 1,456 specimens collected in a small country such as Singapore (730 km^2), we are increasing the number of known Oriental species by ca. 25%. This is an indication for the alarming magnitude of taxonomic impediment and neglect, given that Singapore has lost most of its natural habitats, and the original species diversity was undoubtedly much higher (Brook et al. 2003; Theng et al. 2020). Expanding the geographic scale of a mycetophilid revision to the Malay Peninsula would mean revising the fauna of an area that is 300 times larger (242,363 km^2), with more pristine habitats and elevational gradients that are likely to increase the species diversity. It would surely be impossible to handle the mycetophilid fauna at this scale. One may propose to tackle the fauna one genus at a time, but this would also not be feasible for three species-rich genera that contribute over 60% of the new species (*Neoempheria* Osten-Sacken: N=31; *Epicypta* Winnertz: N=29; *Manota* Williston: N=14).

Working at small geographic scales does, however, require regular combining of information from multiple faunistic revisions covering neighboring areas. For example, our taxonomic revision of Singapore's Mycetophilidae would become essential background information for a taxonomic revision of the same group in additional studies on the fauna of the Malay Peninsula. Fortunately, combining and analyzing barcode data from different studies is straightforward (Meier et al. 2021; Vences et al. 2021) as long as the same barcoding region is sequenced. Such analysis then yields information on the stability and intraspecific variability of each MOTU. This information furthermore guides an expanded LIT analysis involving morphological and/or nuclear data. It is important that this follow-up work should not require the loan of type material to avoid delays and the cost and risks of specimen shipping. This is one major reason why the morphology of new

species should be well documented (Ang et al. 2013). But such documentation is also important for other reasons. It is the only way to provide comparison data for species that were described based on morphology over the last 250 years. This is almost all, given that even now only 10% of all new species descriptions in entomology contain molecular data (Miralles et al. 2021). Integrative species descriptions are also critical for the analysis of specimens not suitable for sequencing and for allowing researchers without access to molecular labs to participate in science. This is a large number of biologists all over the world (Zamani et al. 2022a,b). Neglecting morphology in species description thus leads to the build-up of a new "superficial description impediment" (Meier et al. 2022) in taxonomy because eventually morphological data will have to be collected anyway. Neglect of morphology also seems very shortsighted, given that it appears likely that specimen identification with molecular tools will eventually be complemented with identification based on image-based AI tools (Høye et al. 2021, Wührl et al., 2022). These tools need to be trained with images and arguably the best images come from taxonomic revisions because the specimens have been competently identified (Meier and Dikow 2004).

Establishing integrative species limits with LIT

An important aspect of our study was testing the performance of Hartop et al.´s (2022) Large-Scale-Integrative Taxonomy (LIT). We find that it performs well for Mycetophilidae. Although the initial clustering of barcodes/specimens was at a lower threshold (3%) than was eventually found to maximize congruence between barcode and molecular data, the number of specimens that had to be studied with morphology was moderate (N=225). The proportion of specimens was higher than in the previous application of LIT (Mycetophilidae: 15.5%; Phoridae: 5.1% in Hartop et al. 2022), but this was largely due to the high species/specimen ratio of 8% for Mycetophilidae. Carrying out a LIT analysis for all data (N=3,023) would reduce the proportion of specimens to be studied to 9% (Table 2). In the case of Singapore´s Mycetophilidae, the level of congruence between MOTUs used for specimen sorting (3% clusters) and morphological data was very high (111 congruent MOTUs). This means that only seven morphospecies were in conflict with 3% MOTUs. Note also that the term "conflict" is hardly appropriate from an evolutionary point of view. Measurements such as the match ratio only quantify which proportion of morphospecies are in perfect agreement (i.e., form identical sets) with MOTUs obtained using one particular clustering threshold or algorithm. Yet, the use of a single threshold for clustering barcodes has long been criticized as inappropriate on theoretical grounds (Will and Rubinoff 2004). Indeed, we would expect that closely related species should form several proper subsets within a single MOTU if the latter was delimited using a threshold appropriate for "average species". Conversely, we would expect old species to form union sets composed of several MOTUs. These expectations are in line with the

results of a recent empirical study on optimal sequence similarity thresholds (e.g., Bonin et al. 2023). LIT therefore not only establishes whether morphospecies are "in conflict" with MOTUs, but also tests whether there are clustering thresholds that yield congruent MOTUs. Only if this is not the case, LIT would consider barcode and morphological data to be in genuine conflict and the species would remain undescribed until additional data become available (e.g., nuclear markers, ecology). However, it is comforting that we did not find such a case in our current study or our recent study on Swedish phorids (Hartop et al. 2022).

Publishing dark taxonomy revisions

Many taxonomic monographs consist of two main sections. The introduction and discussion sections contain information that is of interest to a broad readership (e.g., information on biology, species diversity, taxonomic history of a taxon) while the taxonomic section consists of species descriptions that are mostly of interest to taxonomic specialists. Unfortunately, the information that is of general interest is often overlooked by biologists and thus also not cited. We here suggest that a dark taxonomy revision should be published in two separate publications. The taxonomic monograph could only consist of an introduction/method section that presents the background information needed to understand how the species hypotheses were derived, as well as the taxonomic descriptions. Everything else could be in a separate companion paper intended for a broader audience. Such a companion manuscript can, for example, present a rigorous quantitative analysis of the habitat preferences and phenology of species if the revision was based on fresh samples obtained from standardized samples. We believe that it is desirable to bring both manuscripts together in the form of a preprint and this approach is also pursued here. In our general paper, we introduce the concept of dark taxonomy, document that a single, moderately sized revision can increase the number of described species in a dark taxon by 25% for the Oriental region, and present evidence that the species discovered in a first revision cover a very large proportion of specimens collected at different times. In the companion monograph, we describe and illustrate the species and discuss the taxonomic affinity of the newly described fauna (Amorim et al. 2023).

Conclusions

Dark taxa are everywhere. They are abundant, species-rich, and a major obstacle to holistic biomonitoring. What is needed is a robust "dark taxonomy" protocol designed to tackle the taxonomic impediments for this dark diversity. We would argue that whatever protocol may eventually be used should satisfy two criteria: it should be efficient enough to deal with thousands of specimens and species and it should also yield biodiversity baseline data, so that changes in

abundance and richness of dark taxa can be monitored in a not too distant future and utilized for ecosystem management or rehabilitation purposes. We believe that our proposal of "dark taxonomy" can be a major step in this direction. Robots and cost-effective sequencing enable presorting of species to putative species. LIT reduces the amount of morphological work needed to obtain integrative species boundaries. Lastly, barcoding of type specimens will accelerate the process of distinguishing between species that need a scientific name and those that are already described.

Acknowledgements

All work described here was carried out as part of a comprehensive insect survey of Singapore which was carried out in collaboration and with support from the National Parks Board of Singapore (NParks). Special thanks go to Dr. Patrick Grootaert and the team from the National Biodiversity Centre of NParks for their assistance in fieldwork (Permits: NP/RP12-022-4, NP/RP12-022-5, NP/RP12-022-6). We would also like to thank the research staff, lab technicians, undergraduate students, and interns of the Evolutionary Biology Laboratory for their help and assistance. This project would have been impossible without their hard work. Financial support was provided by Tier 2 funding from the Ministry of Education (Singapore).

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(3):478-494.

Amorim D.S., dos Santos C.M.D. 2018. Flies, endemicity, and the Atlantic Forest: a biogeographical study using topographic units of analysis. *Aust. Syst. Bot.* 30:439–469.

Amorim D.S., Oliveira S. S., Balbi M. I. P. A., Ang, Y., Yeo, D., Srivathsan, R., Meier, R. 2023. An integrative taxonomic treatment of the Mycetophilidae (Diptera: Bibionomorpha) from Singapore reveals 115 new species on 730km2. BioRxiv

Ang Y., Wong L.J., Meier R. 2013. Using seemingly unnecessary illustrations to improve the diagnostic usefulness of descriptions in taxonomy–a case study on *Perochaeta orientalis* (Diptera, Sepsidae). *ZooKeys* 355:9–27.

Bechev D. 1995. *Allactoneura papuensis* spec. nov. from New Guinea (Insecta: Diptera: Mycetophilidae). *Reichenbachia* 31:81–82.

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. 2012. The effect of geographical scale of sampling on DNA barcoding. *Syst. Biol.* 61(5): 851–869.

Bickel D.J. 1999. What museum collections reveal about species accumulation, richness, and rarity: an example from the Diptera. In: Ponder W., Lunney, D., editors. *The Other 99%: The Conservation and Biodiversity of Invertebrates*. Royal Zoological Society of New South Wales. p.174–181.

Bickford D., Lohman D.J., Sodhi N.S., Ng P.K.L., Meier R., Winker K., Ingram K.K., Das I. 2007. Cryptic species as a window on diversity and conservation. *Trends Ecol. Ecol.* 22(3):148–155.

Bonin, A., Guerrieri, A., Ficetola, G.F., 2023. Optimal sequence similarity thresholds for clustering of molecular operational taxonomic units in DNA metabarcoding studies. *Molecular Ecology Resources* 23:368–381.

Bonnet X., Shine R., Lourdais O. 2002. Taxonomic chauvinism. *Trends Ecol. Ecol.* 17(1):1–3.

Borkent C.J., Wheeler T.A. 2012. Systematics and Phylogeny of *Leptomorphus* Curtis (Diptera: Mycetophilidae). *Zootaxa* 3549:1–117.

Borkent A., Brown B.V., Adler P.H., Amorim D.S., Barber K., Bickel D., Boucher S., Brooks S.E., Burger J., Burington Z.L., Capellari R.S., Costa D.N.R., Cumming J.M., Curler G., Dick C.W., Epler J.H., Fisher E., Gaimari S.D., Gelhaus J., Grimaldi D.A., Hash J., Hauser M., Hippa H., Ibáñez-Bernal S., Jaschhof M., Kameneva E.P., Kerr P.H., Korneyev V., Korytkowski C.A., Kung G.-A., Kvifte G.M., Lonsdale O., Marshall S.A., Mathis W., Michelsen V., Naglis S., Norrbom A.A., Paiero S., Pape T., Pereira-Colavite A., Pollet M., Rochefort S., Rung A., Runyon J.B., Savage J., Silva V.C., Sinclair B.J., Skevington J.H., Stireman J.O.III, Swann J., Vilkamaa P., Wheeler T., Withworth T., Wong M., Wood D.M., Woodley N., Yau T., Zavortink T.J., Zumbado M.A. 2018. Remarkable fly (Diptera) diversity in a patch of Costa Rican cloud forest: why inventory is a vital science. *Zootaxa* 4402(1):53–90.

Brook B.W., Sodhi N.S., Ng P.K.L. 2003. Catastrophic extinctions follow deforestation in Singapore. *Nature* 424(6947):420–423.

Brown B.V. 2005. Malaise trap catches and the crisis in Neotropical dipterology. *Am. Entomol.* 51(3):180–183.

Brown B.V., Borkent A., Adler P.H., De Souza A.D., Barber K., Bickel D., Boucher S., Brooks S.E., Burger J., Burington Z.L., Capellari R.S., Costa D.N.R., Cumming J.M., Curler G., Dick C.W., Epler J.H., Fisher E., Gaimari S.D., Gelhaus J., Grimaldi D.A., Hash J., Hauser M., Hippa H., Ibáñez-Bernal S., Jaschhof M., Kameneva E.P., Kerr P.H., Korneyev V., Korytkowski C.A., Kung G.-A., Kvifte G.M., Lonsdale O., Marshall S.A., Mathis W., Michelsen V., Naglis S., Norrbom A.A., Paiero S., Pape T., Pereira-Colavite A., Pollet M., Rochefort S., Rung A., Runyon J.B., Savage J., Silva V.C., Sinclair B.J., Skevington J.H.,

Stireman J.O.III., Swann J., Thompson F.C., Vilkamaa P., Wheeler T.A., Witworth T., Wong M., Wood D.M., Woodley N., Yau T., Zavortink T.J., Zumbado M.A. 2018. Comprehensive inventory of true flies (Diptera) at a tropical site. *Commun. Biol.* 1(21): 1–8.

Dayrat B. 2005. Towards integrative taxonomy. *Biol. J. Linn. Soc.* 85(3):407–417.

De Carvalho M.R., Bockmann F.A., Amorim D.S., Brandão C.R.F., de Vivo M., de Figueiredo J.L., Britski H.A., de Pinna M.C.C., Menezes N.A., Marques F.P.L., Papavero N., Cancello E.M., Crisci J.V., McEachran J.D., Schelly R.C., Lundberg J.G., Gill A.C., Britz R., Wheeler Q.D., Stiassny M.L.J., Parenti L.R., Page L.M., Wheeler W.C., Faivovich J., Vari R.P., Grande L., Humphries C.J., DeSalle R., Ebach M.C., Nelson G.J. 2007. Taxonomic Impediment or Impediment to Taxonomy? A Commentary on Systematics and the Cybertaxonomic–Automation Paradigm. *Evol. Biol.* 34:140–143.

Colless D.H. 1966. Diptera: Mycetophilidae. *Insects of Micronesia, Bernice P. Bishop Museum* 12(6):637–666.

Dijkstra K.D.B., Kipping J., Mézière N. 2015. Sixty new dragonfly and damselfly species from Africa (Odonata). *Odonatologica* 44(4):447–678.

Eberhard W.G. 1985. Sexual Selection and Animal Genitalia. Harvard University Press.

Edwards F.W. 1925. British fungus–gnats (Diptera, Mycetophilidae). With a revised generic classification of the family. *Trans. R. Entomol. Soc. Lond.* 73:505–670.

Edwards F.W. 1926. XIV – Diptera Nematocera from the mountains of Borneo. Sarawak Museum Journal III(10):243–278.

Edwards F.W. 1927. Diptera Nematocera from the Dutch East Indies. *Treubia* 9:352–370.

Edwards F.W. 1928. Diptera Nematocera from the Federated Malay States Museum. *Journal of the Federated Malay States Museums* 14:1–139.

Edwards F.W. 1929. Philipinne Nematocerous Diptera III. *Notulae Entomologicae* 9: 70–81.

Edwards F.W. 1931. Fauna Sumatrensis. Mycetophilidae, Diptera. *Beijdrage* 69:262–278.

Edwards F.W. 1933. Diptera Nematocera from Mount Kinabalu. *Journal of the Federated Malay States Museums* 17:223–296.

Edwards F.W. 1935. Two new *Delopsis* collected by Dr. Leefmans in Java. Diptera, Mycetophilidae. *Tijdschrift Voor Entomologie* 78:186–187.

Edwards F.W. 1940. Neotropical Neoempheria (Diptera, Mycetophilidae). *Novitates Zoologicae* 42(1):107–129.

Evenhuis N.L. 2007. Helping solve the “other” taxonomic impediment: completing the eight steps to total enlightenment and taxonomic Nirvana. *Zootaxa* 1407:3–12.

Fernandez-Triana J. L., Whitfield J., Rodriguez J., Smith M., Janzen D., Hajibabaei M., Burns J., Solis A., Brown J., Cardinal S., Goulet H., Hebert P. 2014. Review of *Apanteles* sensu stricto (Hymenoptera, Braconidae, Microgastrinae) from Area de Conservación Guanacaste, northwestern Costa Rica, with keys to all described species from Mesoamerica. *ZooKeys* 383:1–565.

Fernandez-Triana, J.L., 2022. Turbo taxonomy approaches: lessons from the past and recommendations for the future based on the experience with Braconidae (Hymenoptera) parasitoid wasps. *ZooKeys*, 1087:199.

Fitzgerald S.J. 2017. *Phthinia* Winnertz (Diptera: Mycetophilidae): new species and records from the Neotropical and Oriental regions. *Zootaxa* 4231:107–113.

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(5):851 –861.

Hartop E., Srivathsan A., Ronquist F., Meier, R. 2022. Towards large-scale integrative taxonomy (LIT): resolving the data conundrum for dark taxa. *Syst. Biol*. 71(6):1404–1422.

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., Ratnasingham S., Zakharov E.V., Telfer A.C., Levesque-Beaudin V., Milton M.A., Pedersen S., Jannetta P., de Waard J.R. 2016. Counting animal species with DNA barcodes: Canadian insects. *Philos. Trans. R. Soc. Lond. B* 371:20150333.

Hippa H. 2006. Diversity of *Manota* Williston (Diptera: Mycetophilidae) in a Malaysian rainforest: description of twenty-seven new sympatric species. *Zootaxa* 1161:1–49.

Hippa H. 2007. The genus *Manota* Williston (Diptera: Mycetophilidae) in Melanesia and Oceania. *Zootaxa* 1502:1–44.

Hippa H. 2008. New species and new records of *Manota* Williston (Diptera, Mycetophilidae) from the Oriental region. *Zootaxa* 1723:1–41.

Hippa H. 2009. New species and new records of *Manota* Williston (Diptera, Mycetophilidae) from Thailand. *Zootaxa* 2017:1–33.

Hippa H. 2011. New species and new records of *Manota* Williston (Diptera, Mycetophilidae) from Thailand, with a key to the Oriental and Palaearctic species. *Zootaxa* 2763:39–60.

Hippa H., Jaschhof M., Vilkamaa. 2005. Phylogeny of the Manotinae, with a review of *Eumanota* Edwards, *Paramanota* Tuomikoski and *Promanota* Tuomikoski (Diptera, Mycetophilidae). *Studia Dipterologica* 11:405–428.

Hippa H., Kurina O. 2018. Four new species and new records of *Manota* (Diptera: Mycetophilidae) from Sulawesi, Indonesia. *Acta Entomol. Mus. Natl. Pragae* 58:249–256.

Hippa H., Saigusa T. 2016. Notes on Oriental and East Palaearctic *Manota* Williston (Diptera, Mycetophilidae), with the description of seven new species. *Zootaxa* 4084:377–390.

Hippa H., Ševčík J. 2010. Notes on Oriental and Australasian Manotinae (Diptera, Mycetophilidae), with the description of thirteen new species. *Zootaxa* 2333:1–25.

Hippa H., Ševčík J. 2013. Five new species and a new record of *Manota* (Diptera: Mycetophilidae) from Sulawesi. *Acta Entomol. Mus. Natl. Pragae* 53: 763–775.

Høye T.T., Ärje J., Bjerge K., Hansen O.L., Iosifidis A., Leese F., Mann H.M., Meissner K., Melvad C., Raitoharju, J. 2021. Deep learning and computer vision will transform entomology. *Proc. Natl. Acad. Sci. U.S.A.* 118(2):e2002545117.

Huber F., Reis F.H. 2011. Técnica alternativa para montagem de insetos em lâminas permanentes para visualizaçao em microscopia óptica. *EntomoBrasilis* 4:13–19.

Kallweit U. 1998. Notes on the genus *Metanepsia* Edwards and its relatives from East Asia (Insecta: Diptera: Mycetophlidae). *Reichenbachia* 53: 341–353.

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(11):1630–1638.

Karlsson D., Forshage M., Holston K., Ronquist F. 2020. The data of the Swedish Malaise Trap Project, a countrywide inventory of Sweden's insect fauna. *Biodivers. Data J.* 8:e56286

Kaspřák D., Borkent C.J., Wahab R.A. 2017. *Leptomorphus sevciki* sp. nov., a remarkable new wasp-mimicking fungus gnat from Brunei (Diptera: Mycetophilidae). *Acta Entomol. Mus. Natl. Pragae 57*(1):195–203.

Katoh K., Standley D.M. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molec. Biol. Evol.* 30(4):772–780.

Kjaerandsen J., Kerr P., Lindemann J.P., Kurina O. 2023. When details matter: Integrative revision of Holarctic *Coelophthinia* Edwards (Diptera, Mycetophilidae), including mapping of its mitogenome, leads to the description of four new pseudocryptic species. *Biodivers. Data J.* 11: e98741

Kurina O., Hippa H. 2015. A review of the South Pacific *Manota* Williston (Diptera, Mycetophilidae), with description of thirteen new species. *Zootaxa* 4020:257–288.

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.

Laamanen T.R., Petersen F.T., Meier, R. 2003. Kelp flies and species concepts–the case of *Coelopa frigida* (Fabricius, 1805) and *C. nebularum* Aldrich, 1929 (Diptera: Coelopidae). *J. Zool. Syst. Evol. Res.* 41(2):127–136.

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(1):34.

Magnussen T., Soli G.E.E., Kjaerandsen J. 2019. *Allodia* Winnertz from the Himalayas, with nine species new to science (Diptera, Mycetophilidae). *ZooKeys* 820:119–138.

Marsh P., Wild A., Whitfield J. 2013. The Doryctinae (Braconidae) of Costa Rica: genera and species of the tribe Heterospilini. *ZooKeys* 347:1–474.

Matile, L. 1999. Première citation du genre *Parempheriella* Matile dans la Région Paléarctique (Diptera, Mycetophilidae). *Rev. fr. entomol.* 21(3):115–118.

Meier, R., 2017. Citation of taxonomic publications: the why, when, what and what not. *Syst. Ent.*, 42:301–304.

Meier R., Dikow T. 2004. Significance of specimen databases from taxonomic revisions for estimating and mapping the global species diversity of invertebrates and repatriating reliable specimen data. *Conser. Biol.* 18(2):478–488.

Meier R., Kwong S., Vaidya G., Ng P.K.L., 2006. DNA barcoding and taxonomy in Diptera: a tale of high intraspecific variability and low identification success. *Syst. Biol.* 55(5):715–728.

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(1):100-110.

Meier R., Blaimer B.B., Buenaventura E., Hartop E., von Rintelen T., Srivathsan A., Yeo D. 2022. 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(2):264–275. https://doi.org/10.1111/cla.12489

Miralles A., Bruy T., Wolcott K., Scherz M.D., Begerow D., Beszteri B., Bonkowski M., Felden J., Gemeinholzer B., Glaw F., Glöckner F.O. 2020. Repositories for taxonomic data: where we are and what is missing. *Syst. Biol.* 69(6):1231–1253.

Montero□Pau J., Gómez A., Muñoz J. 2008. Application of an inexpensive and high□throughput genomic DNA extraction method for the molecular ecology of zooplanktonic diapausing eggs. *Limnol. Oceanogr.: Methods* 6(6):218–222.

Narakusumo R.P., Balke M., Riedel, A. 2019. Seven New Species of *Trigonopterus* Fauvel (Coleoptera, Curculionidae) from the Tanimbar Archipelago. *ZooKeys* 888: 75 –93.

Narakusumo R.P., Riedel A., Pons J. 2020. Mitochondrial genomes of twelve species of hyperdiverse *Trigonopterus* weevils. *PeerJ* 8:e10017.

Økland B. 1994. Mycetophilidae (Diptera), an insect group vulnerable to forestry practices? A comparison of clearcut, managed and semi-natural spruce forests in southern Norway. *Biodivers. Conserv.* 3: 68 –85.

Økland B. 1996. Unlogged forests: important sites for preserving the diversity of mycetophilids (Diptera: Sciaroidea). *Biol. Conserv.* 76(3):297–310.

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* 371:20150334

Papp L. 2004. Seven new species of Manotinae (Diptera: Mycetophilidae) from Asia and Papua New Guinea. *Acta Zool. Acad. Sci. Hung.* 50:227–244.

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.

Phorid Catalog (http//phorid.net/pcat/flist3.php). Accessed August 2023.

Puillandre N., Lambert A., Brouillet S., Achaz G. 2012. ABGD, Automatic Barcode Gap Discovery for primary species delimitation. *Molec. Ecol.* 21(8):1864–1877.

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.

Riedel A. 2022. Nine new species of *Trigonopterus* Fauvel (Coleoptera, Curculionidae) from Sundaland. *ZooKeys* 1124:109–130.

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

Riedel A., Sagata K., Surbakti S., Tänzler R., Balke M. 2013b. One hundred and one new species of *Trigonopterus* weevils from New Guinea. *ZooKeys* 280:1–150.

Riedel A., Narakusumo R.P. 2019. One hundred and three new species of *Trigonopterus* weevils from Sulawesi. *ZooKeys* 828:1–153.

Riedel A., Tänzler, R. 2016. Revision of the Australian species of the weevil genus *Trigonopterus* Fauvel. *ZooKeys* 556:97–162.

Riedel A., Tänzler R., Balke M., Rahmadi C., Suhardjono Y.R. 2014. Ninety-eight new species of *Trigonopterus* weevils from Sundaland and the Lesser Sunda Islands. *ZooKeys* 467:1–162

Rohner P.T., Ang Y., Lei Z., Puniamoorthy N., Blanckenhorn W.U., Meier, R. 2014. Genetic data confirm the species status of *Sepsis nigripes* Meigen (Diptera: Sepsidae) and adds one species to the Alpine fauna while questioning the synonymy of *Sepsis helvetica* Munari. *Invertebr. Syst.* 28(5):555–563.

Rowe L., Arnqvist G. 2012. Sexual selection and the evolution of genital shape and complexity in water striders. *Evolution* 66(1):40 –54.

Santos D., Ribeiro G.C., Cabral A.D., Sperança M.A. 2018. A non-destructive enzymatic method to extract DNA from arthropod specimens: Implications for morphological and molecular studies. *PLoS One* 13(2):e0192200.

Santos B., Miller M., Miklasevskaja M., McKeown J., Redmond N., Coddington J., Bird J., Miller S., Smith A., Brady S., Buffington M. 2023. Enhancing DNA barcode reference libraries by harvesting terrestrial arthropods at the Smithsonian's National Museum of Natural History. *Biodivers. Data J.* 11:e100904

Savage J., Borkent A., Brodo F, Cumming J.M., Curler G., Currie D.C., deWaard J.R., Gibson J.F., Hauser M., Laplante L., Lonsdale O., Marshall S.A., O'Hara J.E., Sinclair B.J., Skevington J.H., 2019. Diptera of Canada. In: Langor DW, Sheffield C.S., Editors. The Biota of Canada – A Biodiversity Assessment. Part 1: The Terrestrial Arthropods. *ZooKeys* 819: 397–450. https://doi.org/10.3897/zookeys.819.27625

Ševčík J. 2001. First record of *Cordyla* Meigen and *Monoclona* Mik from the Oriental Region (Diptera: Mycetophilidae). *Beitr. Ent.* 51(1):155–160.

Ševčík J., Bechev D., Hippa H. 2011. New Oriental species of Gnoristinae with pectinate antennae (Diptera: Mycetophilidae). *Acta Entomol. Mus. Natl. Pragae* 51(2): 687–696.

Ševčík J., Hippa H. 2010. New species of *Chalastonepsia* and *Pectinepsia* gen. nov. (Diptera: Mycetophilidae) from the Oriental and Australasian Regions. *Acta Entomol. Mus. Natl. Pragae* 50(2): 595–608.

Singapore National Parks Board. 2023. Species List (Red Data Book List). https://www.nparks.gov.sg/biodiversity/wildlife-in-singapore/species-list. Accessed August 2023.

Sivec I., Plassmann E. 1982. Sechs neue Pilzmücken aus Sri Lanka (Diptera, Nematocera, Mycetophilidae). *Spixiana* 5(1):7–13.

Søli G.E.E. 1996. *Chalastonepsia orientalis* gen.n., sp. n., a second genus in the tribe Metanepsiini (Diptera, Mycetophilidae). *Tijdschr. Entomol.* 139:79–83.

Søli G.E.E. 2002. New species of *Eumanota* Edwards, 1933 (Diptera: Mycetophilidae). *Ann. Soc. Entomol. Fr.* 38:45–53.

Srivathsan A., Sha J.C., Vogler A.P., Meier R. 2015. Comparing the effectiveness of metagenomics and metabarcoding for diet analysis of a leaf□feeding monkey (*Pygathrix nemaeus*). *Molec. Ecol. Resourc.* 15(2):250–261.

Srivathsan A., Hartop E., 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. https://doi.org/10.1186/s12915-019-0706-9

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.

Srivathsan A., Ang Y., Heraty J., Hwang W.S., Wan J.F.A., Kutty S.N., Puniamoorthy J., Yeo D., Roslin T., Meier, R. 2023. Convergence of dominance and neglect in flying insect diversity. *Nat. Ecol. Evol.* 7:1012–1021. https://doi.org/10.1038/s41559-023-02066-0.

Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30(9):1312–1313.

Tan D.S., Ang Y., Lim G.S., Ismail M.R.B., Meier, R. 2010. From 'cryptic species' to integrative taxonomy: an iterative process involving DNA sequences, morphology, and behaviour leads to the resurrection of *Sepsis pyrrhosoma* (Sepsidae: Diptera). *Zool. Scr.* 39(1):51–61.

Theng M., Wan J.F.A, Jain A., Huertas B., Tan D.J., Tan H.Z., Kristensen N.P., Meier R., Chisholm R.A. 2020. A comprehensive assessment of diversity loss in a well-documented tropical insect fauna: Almost half of Singapore's butterfly species extirpated in 160 years. *Biol. Conserv.* 242:108401.

Troudet J., Grandcolas P., Blin A., Vignes-Lebbe R., Legendre F. 2017. Taxonomic bias in biodiversity data and societal preferences. *Sci. Rep.* 7(1):1-14.

Van Dam M.H., Laufa R., Riedel A. 2016. Four new species of *Trigonopterus* Fauvel from the island of New Britain (Coleoptera, Curculionidae). *ZooKeys* 582:129 –141.

Vences M., Miralles A., Brouillet S., Ducasse J., Fedosov A., Kharchev V., Kumari S, Patmanidis S., Puillandre N., Scherz M.D., Kostadinov I., Renner S.S. 2021. iTaxoTools 0.1: Kickstarting a specimen-based software toolkit for taxonomists. *Megataxa* 6:77–92. https://www.mapress.com/mt/article/view/megataxa.6.2.1

Walker A.K., Crosby T.K. 1988. The preparation and curation of insects. DSIR, information series 163. Wellington, New Zealand: Science Information Publishing Centre.

Wang W.Y., Srivathsan A., Foo M., Yamane S.K., Meier, R. 2018. Sorting specimen□rich invertebrate samples with cost□effective NGS barcodes: Validating a reverse workflow for specimen processing. *Molec. Ecol. Resourc.* 18(3):490–501.

Wheeler Q.D., Raven P.H., Wilson E.O. 2004. Taxonomy: Impediment or Expedient? *Science* 303(5656):285. DOI: 10.1126/science.303.5656.285

Will K.W., Rubinoff D. 2004. Myth of the molecule: DNA barcodes for species cannot replace morphology for identification and classification. *Cladistics* 20(1):47–55.

Wong W.H., Tay Y.C., Puniamoorthy J., Balke M., Cranston P.S., Meier R., 2014. 'Direct PCR' optimization yields a rapid, cost effective, non-destructive and efficient method for obtaining DNA barcodes without DNA extraction. *Molec. Ecol. Resourc*. 14(6):1271–1280.

World Economic Forum, 2023. The Global Risks Report 2023 (18th ed.). https://www.zurich.com/knowledge/topics/global-risks/the-global-risks-report-2023

Wu H., Yang, C. 1986. A new genus and a new species of Mycomyiinae from Fujian province, China (Diptera, Mycetophilidae). *Acta Entomol. Sin.* 39(1):86–89.

Wu H., Zheng L.-H., Xu H.-C. 2003. First record of *Allodia* Winnertz (Diptera, Mycetophilidae) in China, with description of seven new species. *Acta Zool. Sin.* 28(2):349–355.

Wührl L., Pylatiuk C., Giersch M., Lapp F., von Rintelen T., Balke M., Schmidt S., Cerretti P., Meier R., 2022. DiversityScanner: Robotic handling of small invertebrates with machine learning methods. *Molec. Ecol. Resourc.* 22(4):1626–1638.

Xu H.-C., Wu H. 2002. Two new species of the genus *Azana* from China (Diptera: Mycetophilidae). *Acta Zootax. Sin.* 27:621–623.

Yeo D., Puniamoorthy J., Ngiam R.W.J., Meier R. 2018. Towards holomorphology in entomology: rapid and cost effective adult–larva matching using NGS barcodes. *Syst. Ent.* 43(4):678–691.

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.

Yeo D., Srivathsan A., Puniamoorthy J., Foo M., Grootaert P., Chan L., Guénard B., Damken C., Wahab R.A., Ang Y., Meier R. 2021. Mangroves are an overlooked hotspot of insect diversity despite low plant diversity. *BMC Biol.* 19(1):1–17.

Yu, D. W., Ji, Y., Emerson, B. C., Wang, X., Ye, C., Yang, C., & Ding, Z. 2012. Biodiversity soup: metabarcoding of arthropods for rapid biodiversity assessment and biomonitoring. *Methods Ecol. Evol.*. 3:613–623.

Zaitzev A.I. 1982. Composition and systematic position of the genus *Allactoneura* de Meijere (Diptera, Mycetophilidae). *Entomologicheskoye Obozreniye* 60: 901–913. [In Russian]

Zamani A., Dal Pos D., Fric Z.F., Orfinger A.B., Scherz M.D., Bartoňová A.S., Gante H.F., 2022a. The future of zoological taxonomy is integrative, not minimalist. *Syst. Biodivers*. 20(1):1–14.

Zamani A., Fric Z.F., Gante H.F., Hopkins T., Orfinger A.B., Scherz M.D., Bartoňová A.S., Pos, D.D. 2022b. DNA barcodes on their own are not enough to describe a species. *Syst. Ent.* 47(3):385–389.

Zhang J., Kapli P., Pavlidis P., Stamatakis A. 2013. A general species delimitation method with applications to phylogenetic placements. *Bioinformatics* 29(22):2869–2876.