doi: 10.26492/gbs70(suppl.1).2018-08

Next-Generation identification tools for Nee Soon freshwater swamp forest, Singapore

S.N. Kutty[1], W. Wang[2], Y. Ang[2], Y.C. Tay[1], J.K.I. Ho[1] & R. Meier[1,2]

[1]Evolutionary Biology Laboratory, Department of Biological Sciences,
National University of Singapore,
14 Science Drive 4, 117543 Singapore
meier@nus.edu.sg
[2]Lee Kong Chian Natural History Museum, National University of Singapore,
2 Conservatory Drive, 117377 Singapore

ABSTRACT. Many invertebrate and plant species are difficult to identify even by taxonomic experts. This has created a major obstacle for understanding the ecology of tropical environments. Here we explore the use of new large-scale, cost-effective approaches to species identification using Next-Generation Sequencing ("DNA barcodes"). Due to the rapid drop in sequencing cost, such barcodes have the potential to help with many identification tasks and they will facilitate regular monitoring of habitats. We use this approach to explore the species diversity of Nee Soon freshwater swamp forest and provide taxonomic identification tools for the fauna and flora of the forest. DNA-barcode libraries were generated for the flora (>1000 barcodes; 170 chloroplast genomes) and fauna (ca. 3000 barcodes). In addition, high-resolution images of 502 animal and 200 plant species were placed on an online image database ("Biodiversity of Singapore"). These images are available to help experts and non-experts alike to identify and appreciate these species. The new databases document Nee Soon's impressive diversity, but they are also important for in-depth studies of fauna-floral species interactions. For example, the plant barcodes were used to reconstruct the diet of Raffles' banded langur based on faecal samples. Overall, we find that the fauna in Nee Soon freshwater swamp forest is very diverse and includes many rare species, and that the species composition is very distinct from those living in surrounding habitats. Animal specimens are readily sequenced, while plant specimens (especially those represented by sapwood samples) remain a challenge. However, newer techniques (e.g. based on genome skimming) are starting to help with obtaining plant DNA-barcodes.

Keywords. DNA-barcoding, Next-Generation sequencing, online image database, species discovery

Introduction

One of the most irritating and fascinating properties of tropical environments is that they are species-rich (Ødegaard, 2000). The large number of species means that there are potentially a large number of important biological players in a system. This makes it very difficult to understand critical interactions that sustain the environment. This problem is exacerbated by the fact that the identification of biological specimens to species level is far from straightforward and yet important, because species have different natural histories and species names are used for filing and retrieving biological information (Gotelli, 2004). Because the specimens encountered in ecosystems do not come with species name tags, biologists have to use a variety of techniques for, a) delimiting (grouping specimens into species) (Wheeler & Meier, 2000), b) describing species (Winston, 1999), and c) identifying species (Walter & Winterton, 2007). Not surprisingly, the best techniques for these purposes vary from taxon to taxon. In addition, the best methods for species identification change over time.

From an identification point of view, the most convenient taxa are those wherein species diversity is well understood (i.e., species delimitation and description are quite complete: birds, butterflies), species identification can be accomplished based on readily-accessible features (e.g. morphology, songs, etc.), and the relevant features for identification can be obtained without collecting or even disturbing the animals or plants. Fortunately, many vertebrate species fall into this category. On the other end of the scale are taxa where many/most species are neither delimited or described, and therefore not identifiable. Unfortunately, more than 90% of the world's species (mostly invertebrates) fall into this category (Meier & Dikow, 2004; Ødegaard, 2000). Intermediate along the spectrum of identification feasibility are taxa for which most species have been delimited and described by scientists, but identification is difficult for a variety of reasons. These include the lack of good identification tools (e.g. keys), the reliance on identification features that can only be used by a few taxonomic experts, and the use of identification features that are only visible during certain times of a species' life cycle. Good examples are many insect species that can only be identified based on minute details of genitalia (Ang & Meier, 2010; Pont & Meier, 2002), species of plants that can only be identified when they happen to flower, and insects whose aquatic larval or nymphal stages are unidentifiable because identification tools only exist for adults (e.g. dragonflies, midges; Cranston et al., 2013). It is often the unidentifiable stages that are the most important, from an ecological and biomonitoring point of view (vegetative parts of plants, larval stages of insects).

Identifying most species is a task that can currently only be performed by experts with extensive training in biology. Indeed, for many invertebrate groups there are only a handful of experts worldwide who can identify species (Gotelli, 2004). This unfortunately means that many identification needs of society are not met. An alternative way of identifying species is through the use of so-called "DNA barcodes" (Hebert et al., 2003a; Meier et al., 2006, 2016; Meier, 2008). For animals, most biologists use a small piece of the cytochrome oxidase subunit 1 (*COI*) gene for species identification ("DNA barcode"; Hebert et al., 2003b). This particular gene sequence (barcode) is distinctly different between most species (Kwong et al., 2012a; Hajibabaei et al., 2007; Meier, 2008; but see Kwong et al., 2012b and Meier et al., 2006). One advantage of using DNA barcodes is that it "democratises" the process of species identification. Instead of only having a handful of experts worldwide who can identify species in a particular group, DNA barcodes can be generated by thousands of laboratories around the world. In addition, the cost of obtaining DNA barcodes has been dropping rapidly so that the number of biologists with access to these kinds of data is also increasing rapidly (Wong et al., 2014; Meier et al., 2016). Barcoding all individuals from specimen rich bulk samples with cost-effective high-throughput pipelines also allows for presorting using DNA barcodes and mitigates downstream morphological work on presorted units (Wang et al., 2018). DNA barcodes have the additional advantage of enabling associations between different life history stages (e.g. larvae and adults; see Yeo et al., in press), and the identification of animal and plant parts that are otherwise not diagnostic. For example, DNA fragments can be used to carry out a diet analysis based on DNA remnants in faecal matter (Srivathsan et al., 2015, 2016), while free-floating DNA in water can be used to assess which animals were swimming in the water (Lim et al., 2016).

Advantages aside, the use of DNA barcodes in species identification comes with several caveats that we need to bear in mind; some stem from the nature of species, while others are essentially technical. For example, DNA barcoding uses genes that are not functionally related to the origin of species (Kwong et al., 2012b). Instead, the species-specific signatures in barcode genes are due to the fact that most species pairs are old enough that sister species are distinguishable based on the genetic differences that accumulated over evolutionary time through a mixture of genetic drift and natural selection (Meier, 2008). Predictably, recently diverged species pairs can share DNA barcodes; i.e., they cannot be distinguished based on these barcodes. Based on ten years of experience with barcoding, this is fairly rare in animal species and about 90% of all species have their own signature in *COI* sequences. A bigger problem that is more technical in nature is that a large proportion of animal species are not yet barcoded which interferes with the use of DNA barcodes for species identification (Kwong et al., 2012a). This is unfortunate because many environmental problems can be diagnosed using DNA barcodes, e.g. the presence of invasive species (Collins et al., 2012; Ng et al., 2016). As for plants, their genes evolve slower so that there is a larger proportion of closely-related species that are indistinguishable based on DNA barcodes (Hollingsworth, 2008; Hollingsworth et al., 2011). This means that DNA barcodes can often only distinguish plant genera. One solution to this problem – which was also pursued in this study – is sequencing multiple genes or whole chloroplast genomes (see "genome skimming"; Straub et al., 2012).

Other technical problems with DNA barcodes are mostly related to cost and time. In particular, traditional Sanger-based DNA barcodes are very expensive (consumables and manpower). Fortunately, we recently developed Next-Generation-Sequencing (NGS)-based DNA barcodes that circumvent these problems (Meier et al., 2016). This is why we were also able to barcode a large number of Nee Soon specimens and use this information for species discovery. Another technical problem is the large amount of Polymerase Chain Reaction (PCR)-inhibitors in DNA extracts of plants. This interferes with amplifying plant barcodes. We addressed this issue through the use of different extraction techniques and by using genome skimming for obtaining chloroplast genomes. The latter has fewer amplification problems and yields more data at roughly the same cost because the cost per base pair of DNA is much lower for NGS than Sanger sequencing.

Identification of specimens via DNA barcodes is slower than identification via morphology for those species with obvious diagnostic morphological features. For

Sanger barcodes, the normal time between collection and obtaining identification is two working days while it can be several weeks for cost-effective barcoding via NGS. This is why morphology is the identification technique of choice for all species where the relevant morphological features are obvious and easily accessible (or can be made assessable through better imaging). Such data can now be conveniently displayed online in digital reference collections (Ang et al., 2013a) and modern publishing also allows for image-rich species descriptions (Ang et al., 2013b). Ideally, morphology and DNA should be combined in recognising species and providing identification tools (Tan et al., 2010; Rohner et al., 2014). Such "integrative taxonomy" is most likely to identify accurate species boundaries and allows for species identification based on either type of data.

In the Nee Soon hydrology and biodiversity project (Clews et al., 2018; Davison et al., 2018), our team used a wide variety of techniques to tackle species identification problems and to create tools for the future. The ultimate goal was to enable and to make it easier to identify biological specimens from Nee Soon freshwater swamp forest. A secondary goal was to generate more "democratic" identification tools; i.e., to provide tools that are less reliant on expensive and rarely available taxonomic expertise. Democratisation of species identification can be achieved by generating higher-quality images that help non-experts identify species (Ang et al., 2013a). This approach was pursued for many animal and plant taxa and we generated a species database in which more than 500 species are illustrated. This database is also a colourful celebration of Nee Soon freshwater swamp forest's glorious biodiversity.

Main objectives

1) Insect diversity. In order to explore the insect diversity of Nee Soon freshwater swamp forest, we used NGS-based species discovery techniques for targeting taxa that belong to different ecological guilds.

2) Faunal identifications: Nee Soon freshwater swamp forest was extensively surveyed by Faunal Ecology teams (Ho et al., 2018; Lim et al., 2018) who collected a large number of specimens. We imaged these specimens using specialised digital camera systems and sequenced the *COI* barcode for these specimens using Sanger sequencing and NGS.

3) Floral identifications: The flora of Nee Soon freshwater swamp forest was studied by the Vegetation Ecology team (Chong et al., 2018) who provided samples for DNA barcoding. Initially, we targeted multiple plant barcode genes (*matK*, *rbcL*, *trnL*, etc.) but due to PCR-inhibitors this was very expensive in terms of manpower and consumables. We therefore switched to NGS-based sequencing of chloroplast genomes via genome skimming. The sequencing efforts concentrated on trees and lianas because they are most relevant for understanding the vegetation ecology of Nee Soon freshwater swamp forest. To assist the floral field team with identifying tree

species in which taxonomically important parts were not readily accessible, we also developed a NGS-based technique for identifying trees to genus based on sapwood samples. This was challenging because such samples contain little DNA and large amounts of PCR-inhibitors.

Methods

Faunal barcoding

Barcodes for insects were generated through *COI* amplification using direct-PCR (dPCR) (Wong et al., 2014), where a small amount of tissue is dissected from each individual specimen and serves as a template for amplification without prior DNA extraction. In addition, DNA from many specimens was also extracted using a novel reagent known as QuickExtract^{TM} DNA Extraction Solution (EPICENTRE Biotechnologies); DNA extracts obtained with QuickExtract were used directly as input template for amplification of barcoding genes. A short fragment (313 bp) of *COI* was used as the general faunal barcoding gene. Subsequent downstream sequencing was conducted using a combination of traditional Sanger sequencing and high-throughput, paired-end NGS on Illumina^{TM} platforms. Most barcodes were generated with NGS (Srivathsan et al., 2015). For NGS barcodes, each specimen's amplicons were tagged with a uniquely labelled primer pair in the PCR step; the use of indexed primers allowed for barcodes to be traced accurately to their specimen of origin in the downstream bioinformatic process. Sequences generated from either Sanger or NGS methods were aligned using MAFFT ver.7 (Katoh & Standley, 2013), before being grouped into molecular operational taxonomic units (mOTUs) based on objective clustering, whereby sequences are grouped by similarity based on uncorrected pairwise (p) distances at specific uncorrected percentage thresholds (Meier et al., 2006; Srivathsan & Meier, 2012). Ideally, the threshold value set for clustering into mOTUs should be within a numerical range where the number of clusters remains stable; this is because stable clusters are likely to represent species. When used appropriately, intraspecific and interspecific variability can also be compared to further assess the stability of the species boundaries (Meier et al., 2008).

Floral barcoding

DNA extraction for all floral specimens were carried out on leaf and sapwood samples using a modified CTAB-chloroform extraction protocol (Doyle & Doyle, 1987). DNA barcodes for plant species were generated using both Sanger sequencing and NGS platforms. Initially, sequences for four selected barcode genes were amplified and sequenced: *maturase K* (*matK*), *ribulose-bisphosphate carboxylase* (*rbcL*), a chloroplast *tRNA* gene (*trnL*), and the *internal transcribed spacer 2* (*ITS2*) region of nuclear ribosomal DNA. The *trnH–psbA* intergenic spacer region was also explored.

With the advent and availability of high throughput sequencing, Illumina^{TM} platforms were used for both amplicon sequencing and chloroplast genome skimming.

Using tagged primers similar to the protocol used for *COI* barcoding, data was generated on a MiSeq sequencer. To cost-efficiently carry out genome skimming for ~240 species, multiplexing was done by ligating a 20bp species-specific tag to the DNA of different species using a modified version of the Meyer & Kircher (2010) protocol. Three such "plant pools" libraries comprising of 75–100 species each with insert sizes of 400–900bp were prepared and sequenced on a HiSeq 2500 (250PE) platform. The sequence data were demultiplexed based on species-specific tags using SABRE (https://github.com/najoshi/sabre) and quality checked using custom scripts. Trimming of reads were carried out in CLC Genomics Workbench (Limit=0.001, https://www.qiagenbioinformatics.com) and assembled into chloroplast contigs using default parameters in MITOBIM (Hahn et al., 2013), by iterative mapping onto a closely-related reference chloroplast genome. Species reads were mapped to the MITOBIM-assembled contigs in CLC Genomics Workbench to calculate the average coverage of each chloroplast genome.

Tree identifications via sapwood samples

The suitability of all four plant barcodes for tree identification via sapwood material was assessed in preliminary experiments, which showed the ~400bp fragment of the *ITS2* marker to be most effective at identification. However, PCR amplification and Sanger sequencing successes with this marker were low due to length variants (success rates of only ~30%). Hence, we switched to using the *trnL* markers (short fragment of 10-50bp) for sapwood-based identifications of the remaining samples unidentifiable with the previous marker. Between one to five PCR amplifications using tagged primers for the *trnL* marker were performed for each sample, and sequenced on an Illumina^{TM} MiSeq Nano run. Sequence data obtained from the run were demultiplexed and binned into unique read clusters using PEAR (Zhang et al., 2014) and OBITOOLS (Boyer et al., 2016) respectively. Consensus sequences of each unique cluster were then matched against both the global and local plant *trnL* databases for identifications via blastn (BLAST 2.2.28+, Camacho *et al.*, 2009).

Specimen imaging and online database

Photography and image preparation. The specimens are kept at the Lee Kong Chian Natural History museum (specimens in main collection and DNA in cryo-collection). One specimen per species was imaged using a high-resolution photomacrography system (Visionary Digital^{TM} Lab Plus System). Specimens were imaged under high magnification at different focal depths and exported via Adobe Lightroom. These images were then digitally stacked into a completely focused composite image using Helicon Focus Pro. The composite images were then digitally optimised in Photoshop CS5 Extended by white-balancing, image sharpening, light/shadow adjustments, and digitally removing impurities from background and specimens. Depending on the taxon, specimens were imaged in different orientations and magnifications to illustrate key diagnostic features. These separate images were then digitally stitched into an image plate.

Image plates were exported in a format that allows for online magnification. This allows the users on internet browsers to view both an overall view of the plate but also zoom in to high-magnification images of structures that are critical for identification: it divides an image into a series of smaller-sized picture tiles at different resolutions and sizes that are presented onto a fixed frame. Because the viewer frame requires only few picture tiles to be loaded at any time, viewing is fast and smooth.

Online image database. All images are displayed on an online image database (https://singapore.biodiversity.online/). The website also displays other species from Singapore. The webpage has a collapsible taxon-based navigation panel on the left, while the main field displays all imaged specimens (in thumbnails). There are also filter options. The image database features a three-tiered design; users can select a taxon group on the left navigation panel, which will show all the available species on the right panel, segregated by another two taxon levels (usually, order, followed by family or genus). Clicking on a species thumbnail will direct users to the species page, which displays the image plate for the specimen, as well as other basic information such as species name, common name, taxonomic information, image information as well as other additional species information and links to other websites for more information on the species (where available). Users can then navigate to other taxon groups using navigation panels.

Results

Faunal barcoding

Using a combination of traditional Sanger sequencing and high throughput Next Generation Sequencing (NGS barcoding, we generated 2904 animal barcodes (predominantly *COI*, with some *COII* for Odonata) (Table 1). Overall, we have a total of 2904 barcodes from the faunal specimens collected from Nee Soon freshwater swamp forest. The majority of barcodes are for insects: Diptera (1399), ants (652) and Odonata (347). Insects collected from Nee Soon's waterways are represented by 128 barcodes. Fishes represent the next largest contribution (201) of barcodes for non-insect fauna. However, numerous additional barcodes continue to be generated for the material that was collected. All barcodes were checked against Genbank in order to rule out contamination.

DNA barcode database for plants

A total of 1189 barcodes were generated for 503 species representing 98 plant families, using a combination of traditional Sanger sequencing and high throughput Next Generation Sequencing. We also generated genome data for 240 species from which 170 chloroplast genomes could be assembled via genome skimming. Coverage of chloroplast genomes depended on both sequencing depth and amount of chloroplast

Table 1. Faunal barcodes generated for Nee Soon.

Taxa	**Overall no. of barcodes**
Fishes	**201**
Mollusca (Snails)	**4**
Crabs and Shrimp (Decapoda)	**12**
Damsel- and Dragonflies (Odonata)	**347**
Bees (Anthophila)	**49**
Ants (Formicidae)	**652**
Termites (Isoptera)	**112**
Fungus Gnats (Mycetophilidae)	**875**
Mosquitoes (Culicidae)	**320**
Horse Flies (Tabanidae)	**5**
Hover Flies (Syrphidae)	**8**
Soldier Flies (Stratiomyidae)	**19**
Chironomidae (Non-biting midges)	**170**
Ceratopogonidae (Biting Midges)	**2**
Baetidae (Small Minnow Mayflies)	**3**
Caenidae (Squaregill Mayflies)	**3**
Gerridae (Water Striders)	**1**
Nepidae (likely *Ranatra*)	**1**
Hemipteran (likely leafhopper)	**1**
Gyrinidae (Diving beetles)	**1**
Scirtidae (Marsh beetles)	**1**
Nemouridae (Stoneflies)	**3**
Perlidae (Stoneflies)	**2**
Calamoceratidae (Caddisfly)	**2**
Dipseudopsidae (Caddisfly)	**9**
Ecnomidae (Caddisfly)	**10**
Hydropsychidae (Caddisfly)	**32**
Hydroptilidae (Caddisfly)	**6**
Leptoceridae (Caddisfly)	**26**
Polycentropodidae (Caddisfly)	**24**
Psychomyiidae (Caddisfly)	**2**
Blattodea (Cockroach)	**1**
	Total: 2904

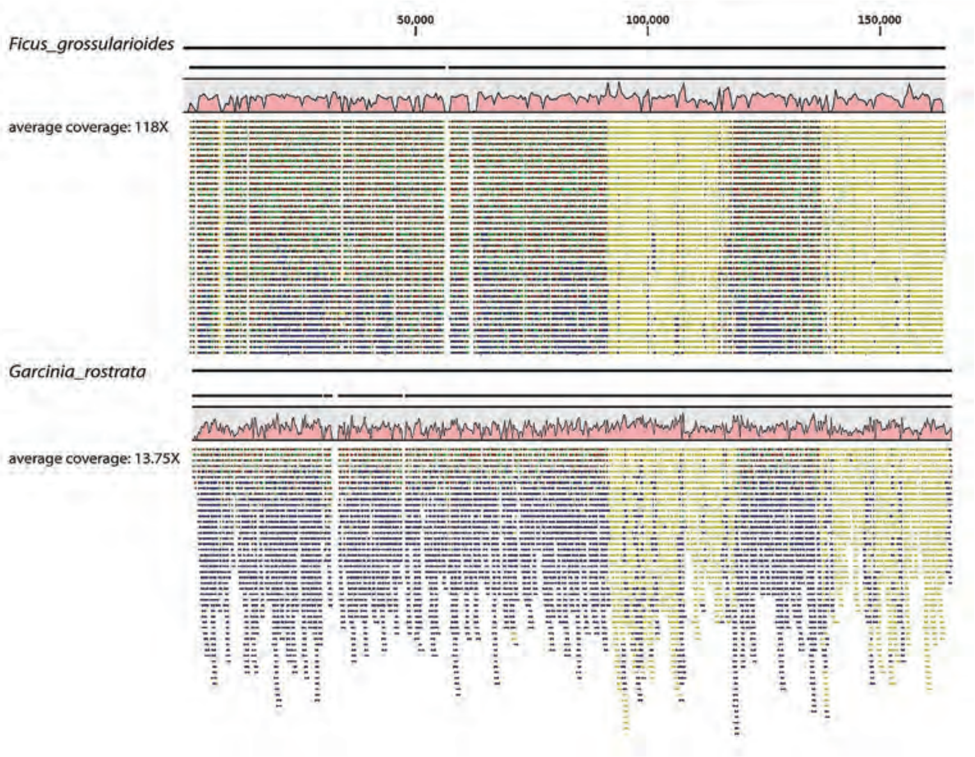


Fig. 1. Variation in coverage of chloroplast genomes (>150,000 bp) between representative Nee Soon species.

material in each species. Based on our data, chloroplast reads varied from 0.23% to 12% of total reads depending on sample. These factors contributed to the uneven genome coverage (see Fig. 1).

Tree identifications via sapwood samples

A total of 360 amplicons from inner bark and sapwood scrapes were sequenced. Based on the local *ITS2* database generated from the leaf samples, Sanger DNA barcodes allowed identification of 40 of the 87 sapwood samples with high confidence to at least the genus level (>=95% sequence match), and 8 samples to at least the family level (90-95% sequence match). Of the 173 sapwood samples that were sent for Illumina sequencing of the short *trnL* gene fragment, 56 and 29 samples were identified with high confidence to the family and genus levels respectively. However, we also find molecular identifications that conflict with expected IDs that were obtained with morphological means; i.e., the technique requires more testing.

Specimen imaging and databasing

A total of 502 faunal specimens originating from Nee Soon freshwater swamp forest were imaged and uploaded onto the Biodiversity of Singapore database: (Fig.

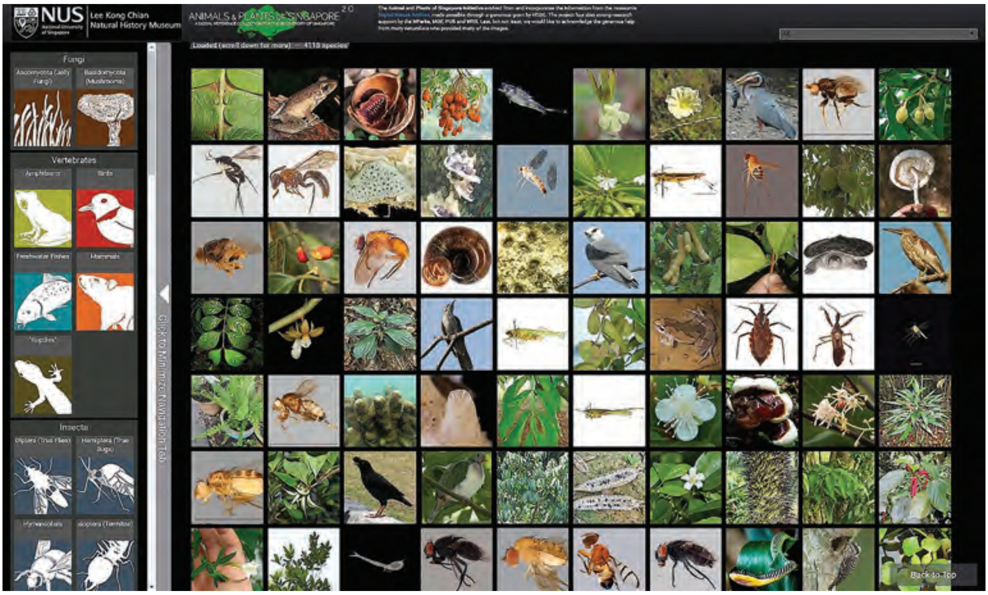

NUS
National University of Singapore
Lee Kong Chian Natural History Museum
ANIMALS & PLANTS of SINGAPORE 2.0
Loaded (scroll down for more) — 4110 species
Fungi
Ascomycota (Jelly Fungi)
Basidiomycota (Mushrooms)
Vertebrates
Amphibians
Birds
Freshwater Fishes
Mammals
"Reptiles"
Insecta
Diptera (True Flies)
Hemiptera (True Bugs)
Hymenoptera
Isoptera (Termites)
Click to Minimize Navigation Tab
Back to Top

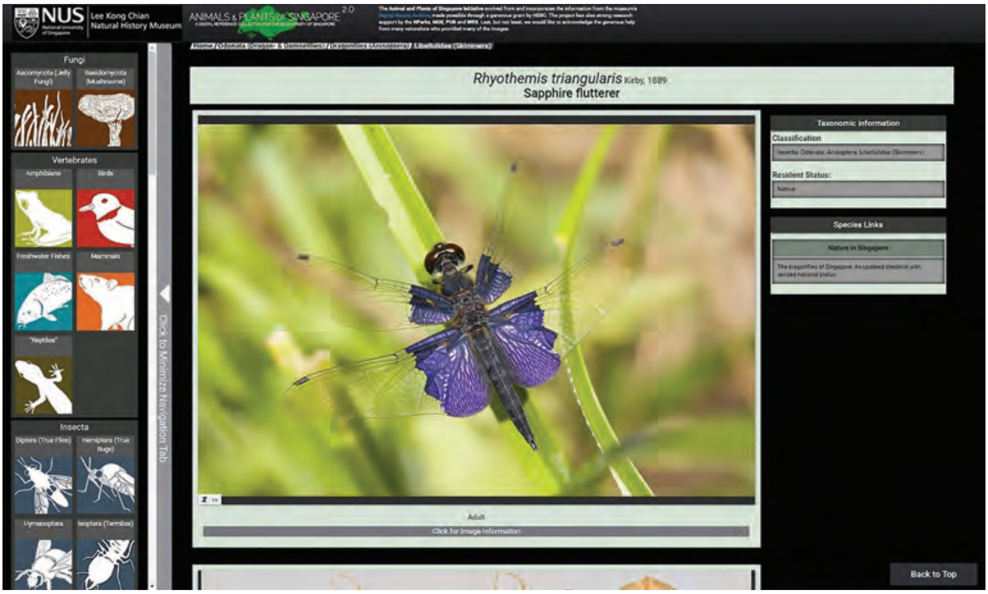
Fig. 2. The home page of the Biodiversity of Singapore website.

2: https://singapore.biodiversity.online/) as individual species pages (Fig. 3). This includes specimens from multiple groups, including Vertebrata, Crustacea, Mollusca, Coleoptera, Odonata, Blattodea, Ephemeroptera, Trichoptera, Plecoptera, Hemiptera and Diptera (Fig. 4). Additionally, images for more than 200 plant species were also added into the database. Table 2 shows the breakdown by taxa. These high resolution images allow for a close-up view of the finer details of the individual species as well (Fig. 5).

Discussion

We set out to use a wide variety of techniques to tackle species identification problems and to create identification tools for the future. As documented in the Results section, we succeeded to various degrees. In addition, the material that was collected during the project continues to be studied and new species are found and imaged every week.

The Faunal Ecology teams extensively surveyed the aquatic habitats and collected a large number of specimens (Ho et al, 2018; Lim et al., 2018). Because the processing of these samples was time-consuming, we only obtained them fairly late. Nevertheless, our preliminary results indicate that the species diversity of Nee Soon is very high and that most species found in the aquatic environments of Nee Soon freshwater swamp forest are distinctly different from what is found in the nearby reservoirs. A good example are the chironomid midges where preliminary sampling revealed more than 250 species in Nee Soon, while only about 40 species were found



Rhyothemis triangularis Kirby, 1889
Sapphire flutterer

Taxonomic information
Classification: Insecta: Odonata: Anisoptera: Libellulidae (Skimmers)
Resident Status: Native

Species Links
Nature in Singapore
The dragonflies of Singapore: An updated checklist with revised national status

Fig. 3. An example of a species page from the Biodiversity of Singapore website.

in the surrounding reservoirs – with some species reaching nuisance levels (Cranston et al., 2013). What is remarkable is that only very few species are shared (unpublished data). We see similar results emerging for other taxa (Odonata, Trichoptera). However, a full evaluation will take some time.

We have also carried out species discovery projects on terrestrial insect groups including ants, termites, fungus gnats (indicators of fungal diversity), stratiomyids (soldier flies), syrphids (hover flies that are pollinators), and mosquitoes. Most of these taxa are very species rich in Nee Soon and the fauna is again very distinct from other habitats for which data exists (NUS campus, mangrove fragments in Singapore). Particularly remarkable is the very high species diversity in fungus gnats (over 200 species). In order to explore the diversity in so many insect groups, we relied on DNA barcodes. To date, we have generated more than 3000 barcodes for the fauna of Nee Soon freshwater swamp forest. The DNA barcodes for different specimens were compared and grouped based on overall similarity (3% threshold) (Meier et al., 2008; Srivathsan & Meier, 2012). Afterwards, one specimen for each cluster was selected for imaging. So far, over 500 images of Nee Soon species have been added to the publicly accessible database "Biodiversity of Singapore", while the entire database includes over 10000 species. However, many additional species are being added every month. Important recent additions are images for the *c*. 200 species of fungus gnats. Most of the species are currently only known from Nee Soon freshwater swamp forest and many species are only known from a single specimen. Such rarity is a pervasive pattern of many tropical species (Lim et al., 2011).

We also succeeded in barcoding most of the important plant species in Nee Soon

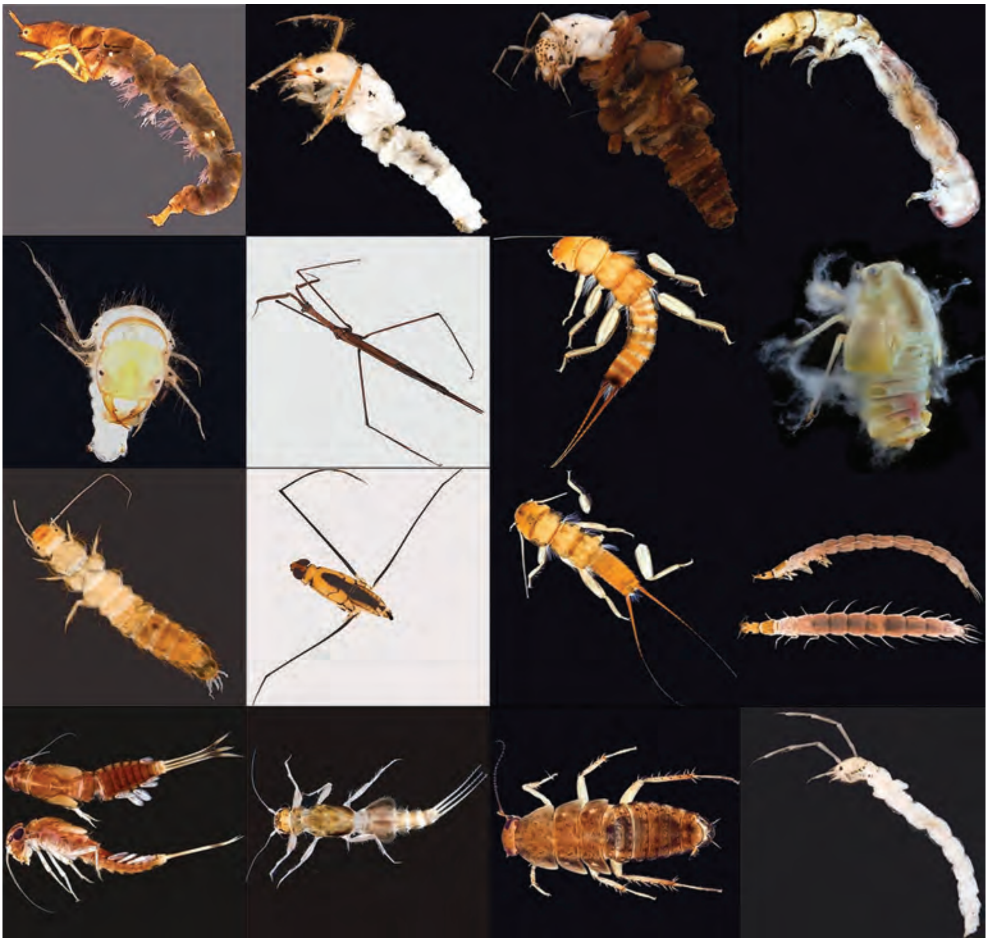


Fig. 4. Some of the diversity of macroinvertebrates found in Nee Soon.

(total of 1189 barcodes). The barcoding gene with the highest sequencing success was *rbcL* (321 species). Unfortunately, this gene rarely allows for distinguishing closely related species. We therefore invested more resources and time into getting sequences for *matK* (275 spp), *trnL* (321 species), *ITS2* (190 spp), and *TrnH-psbA* (86 spp). The value of these barcodes was immediately illustrated when they were used to analyse the diet of Singapore's critically endangered Raffles' banded langur population (Ang et al., 2010; Ang et al., 2012) based on their faecal samples (Srivathsan et al., 2015, 2016). The faecal material included DNA signatures for more than 50 plant species (Ang et al., 2013a) and it will be important for the conservation of Raffles banded langur to keep healthy populations of food plants. Unfortunately, obtaining these plant barcodes via PCR and Sanger sequencing was extremely time-consuming, therefore we also developed a different approach via genome skimming (low coverage sequencing of whole-genomes). Based on the results, we predict that genome skimming will be

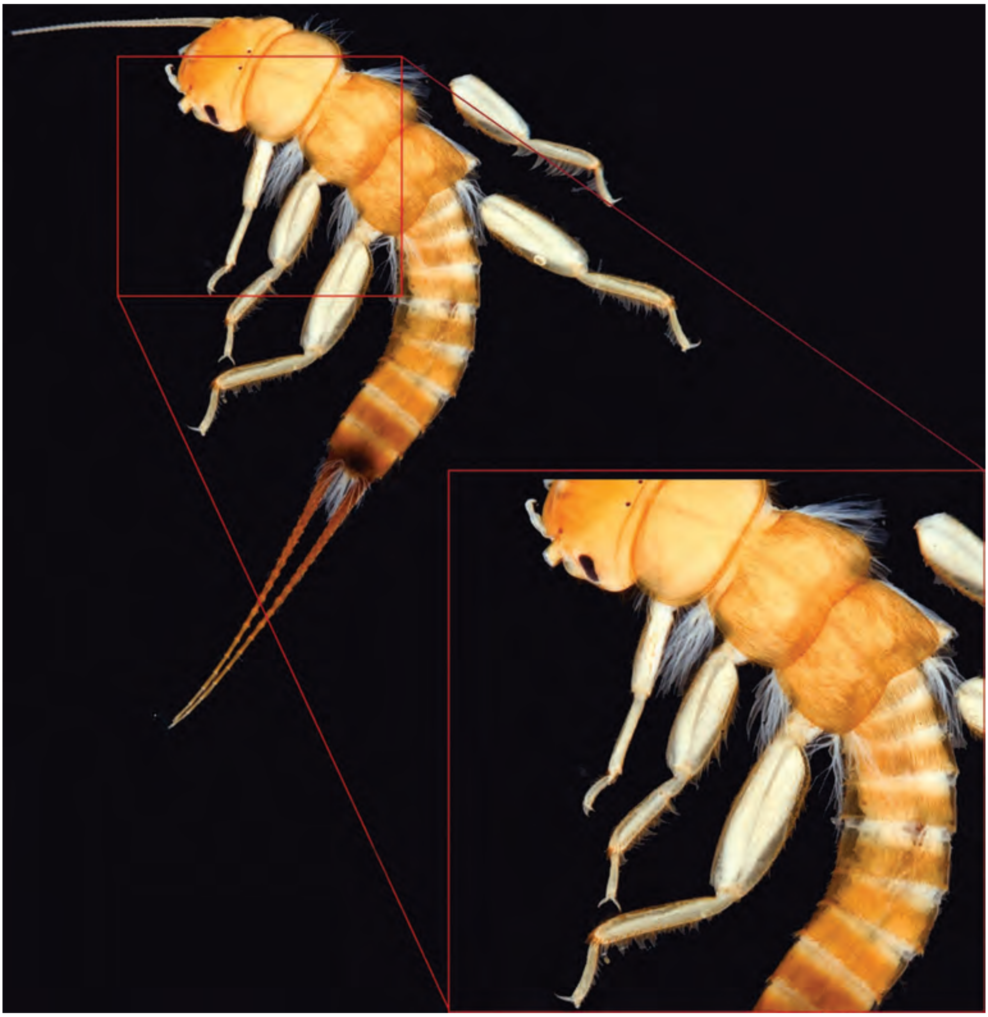


Fig. 5. A close-up view of one of the species found in Nee Soon.

the future technique of choice because it is not only cost-effective, but also yields more data. Genome skimming is a viable alternative to plant barcoding, because vegetative tissues are naturally enriched with chloroplast genes, thus low coverage sequencing of whole-genome extractions can yield enough data for obtaining chloroplast genomes. Whole chloroplast genomes automatically cover most plant barcoding genes that are located on this genome, while yielding much more information than barcoding genes, because whole genomes are much longer (150,000 bp) than all barcoding genes combined (<2000 bp). The work yielded chloroplast genomes for ~170 species, but we hope to eventually cover much of Singapore's flora. A barcode database for all species would contribute towards understanding species interactions as illustrated by our work on the diet of Raffles' banded langur.

Table 2. Breakdown for images of species according to taxa featured on the image database.

Taxon group	**No. species/MOTUs featured**
Vertebrates	**(subtotal: 97)**
Fishes	53
Anura (Frogs)	16
Aves (Birds)	18
Mammalia (Mammals)	10
Crustacea	**(subtotal: 9)**
Decapoda (Shrimps)	5
Brachyura (Crabs)	4
Mollusca	**(subtotal: 7)**
Gastropoda (Terrestrial snails)	7
Diptera (True Flies)	**(subtotal: 138)**
Dolichopodidae (Long-legged Flies)	15
Chironomidae (Non-biting Midges)	1
Culicidae (Mosquitoes)	35
Mycetophilidae (Fungus Gnats)	78
Stratiomyidae (Soldier Flies)	7
Ceratopogonidae (Biting Midges)	2
Odonata	**(subtotal: 35)**
Anisoptera (Dragonflies)	19
Zygoptera (Damselflies)	16
Blattodea	**(subtotal: 33)**
Isoptera (Termites)	32
Cockroach	1
Ephemeroptera (Mayflies)	**(subtotal: 2)**
Baetidae (Small Minnow Mayflies)	1
Caenidae (Small Squaregill Mayflies)	1

Table 2. Continuation.

Taxon group	No. species/MOTUs featured
Trichoptera (Caddisflies)	**(subtotal: 7)**
Calamoceratidae	2
Ecnomidae	1
Hydropsychidae	1
Leptoceridae	3
Plecoptera (Stoneflies)	**(subtotal: 2)**
Perlidae	2
Hemiptera (True Bugs)	**(subtotal: 3)**
Gerridae (Pond Skaters)	1
Nepidae (Water Scorpions)	1
Cicadomorpha (likely leafhopper)	1
Coleoptera (Beetles)	**(subtotal: 2)**
Gyrinidae (Whirligig beetles)	1
Scirtidae	1
Plants	**(subtotal: 164)**
Peridophyta (Ferns)	2
Monocots	5
Magnoliids	45
Rosids	78
Asterids	22
‘other’ Eudicots	12
	Total: 502

Lastly, we used Sanger barcodes and developed NGS barcodes for identifying trees to genus based on sapwood samples. This is often needed because it is difficult to obtain leaves from tall trees; frequently, the only available material is sapwood sampled from the tree trunk. Unfortunately, obtaining DNA from such samples is challenging because the DNA content is small and the DNA extractions include large amounts of PCR-inhibitors. In order to succeed, we first built an *ITS2* database based on leaf samples and used Sanger barcodes to identify most of the 89 sapwood samples to at least family level. However, we were not able to use the same approach for the remaining sapwood samples because there were too many problems with DNA amplification and sequencing. We therefore switched to NGS barcoding of a short *trnL* gene fragment for 169 sapwood samples. The advantage of this approach is that the shorter fragments are more likely to amplify, but this comes at the cost of these fragments containing less information. Using this approach, we were able to identify 44, and 29 samples with high confidence to the family and genus levels respectively. Compared to barcoding of leaf samples, sapwood samples will remain very problematic, hence new approaches should continue to be pursued. Particularly promising may be anchored hybrid enrichment of chloroplast genes.

Conclusions

Being able to identify specimens to species is important for most in-depth study of biological systems. However, obtaining these identifications is very challenging in tropical environments. Fortunately, a number of new tools make this task less daunting. New imaging techniques help with illustrating relevant characters and new and cheaper DNA barcodes allow for the generation of databases that can be used by many researchers. Overall, making the fauna and flora of Nee Soon freshwater swamp forest and Singapore identifiable is achievable. Several hundred, or even thousands of species may potentially be revealed from samples that have been collected and stored. By focusing on particular taxa belonging to different ecological guilds, it is feasible to start understanding species turnover rates across habitats in Singapore and to use this information for conserving Singapore's native fauna and flora. A particularly high priority is obtaining plant barcodes for all of Singapore's vascular plant species. This will allow for in-depth studies of species interactions between plants and animals (e.g. pollination).

ACKNOWLEDGEMENTS. This study forms part of the research project "Nee Soon Swamp Forest Biodiversity and Hydrology Baseline Studies—Phase 2" funded by the National Parks Board (NParks), Singapore. The authors are also grateful to the Public Utilities Board (PUB), Singapore for making this study possible.

References

Ang, Y. & Meier, R. (2010). Five additions to the list of Sepsidae (Diptera) for Vietnam: *Perochaeta cuirassa* sp. n., *Perochaeta lobo* sp. n., *Sepsis spura* sp. n., *Sepsis sepsi* Ozerov, 2003 and *Sepsis monostigma* Thompson, 1869. *Zookeys* 70: 41–56.

Ang, A., Ismail, M.R.B. & Meier, R. (2010). Reproduction and infant pelage colouration of the banded leaf monkey (Mammalia: Primates: Cercopithecidae) in Singapore. *Raffles B. Zool.* 58: 411–415.

Ang, A., Srivasthan, A., Md-Zain, B.M., Ismail, M.R.B. & Meier, R. (2012). Low genetic variability in the recovering urban banded leaf monkey population of Singapore. *Raffles B. Zool.* 60(2): 589–594.

Ang Y., Puniamoorthy, J., Pont, A.C., Bartak, M., Blanckenhorn, W.U., Eberhard, W.G., Puniamoorthy, N., Silva, V.C., Munari, L. & Meier, R. (2013a). A plea for digital reference collections and other science-based digitization initiatives in taxonomy: Sepsidnet as exemplar. *Syst. Entomol.* 38: 637–644.

Ang, Y., Wong, L.J. & Meier, R. (2013b). 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.

Boyer, F., Mercier, C., Bonin, A., Le Bras, Y., Taberlet, P. & Coissac, E. (2016). OBITOOLS: a UNIX-inspired software package for DNA metabarcoding. *Mol. Ecol. Resour.* 16: 176–182.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K. & Madden, T.L. (2009). BLAST+: Architecture and applications. *BMC Bioinformatics* 10: 421.

Chong, K.Y., Lim, R.C.J., Loh, J.W., Neo, L., Seah, W.W., Tan, S.Y. & Tan, H.T.W. (2018). Rediscoveries, new records, and the floristic value of the Nee Soon freshwater swamp forest, Singapore. *Gard. Bull. Singapore* 70 (Suppl. 1): 49–69.

Clews, E., Corlett, R.T., Ho, J.K.I., Kim, D.E., Koh, C.Y., Liong, S.Y., Meier, R., Memory, A., Ramchunder, S.J., Sin, T.M., Siow, H.J.M.P., Sun, Y., Tan, H.H., Tan, S.Y., Tan, H.T.W., Theng, M.T.Y., Wasson, R.J., & Yeo, D.C.J. & Ziegler, A.D. (2018). The biological, ecological and conservation significance of freshwater swamp forest in Singapore. *Gard. Bull. Singapore* 70 (Suppl. 1): 9–31.

Collins, R.A., Armstrong, K.F., Meier, R., Yi, Y., Brown, S.D.J., Cruickshank, R.H., Keeling, S. & Johnston, C. (2012). Barcoding and border biosecurity: identifying cyprinid fishes in the aquarium trade. *PLoS ONE* 7(1): e28381.

Cranston, P., Ang, Y., Heyzer, A., Lim, R.B.H., Wong, W.H., Woodford, J.M. & Meier, R. (2013). The nuisance midges (Diptera: Chironomidae) of Singapore's Pandan and Bedok reservoirs. *Raffles B. Zool.* 61: 779–793.

Davison, G.W.H., Cai, Y.X., Li, T.J. & Lim, W.H. (2018). Integrated research, conservation and management of Nee Soon freshwater swamp forest: hydrology and biodiversity. *Gard. Bull. Singapore* 70 (Suppl. 1): 1–7.

Doyle, J.J. & Doyle, J.L. (1987). A rapid DNA isolation procedure for small quantities of fresh leaf tissue. *Phytochem. Bull.*19: 11–15.

Gotelli, N.J. (2004). A taxonomic wish-list for community ecology. *Philos. T. Roy. Soc. B* 359: 585–597.

Hahn, C., Bachmann, L. & Chevreux, B. (2013). Reconstructing mitochondrial genomes directly from genomic next-generation sequencing reads—a baiting and iterative mapping approach. *Nucleic Acids Res.* 41: e129.

Hajibabaei, M., Singer, G.A.C., Hebert, P.D.N. & Hickey, D.A. (2007). DNA barcoding: how it complements taxonomy, molecular phylogenetics and population genetics. *Trends Genet.* 23 (4): 167–172.

Hebert, P.D.N., Cywinska, A., Ball, S.L. & deWaard, J.R. (2003a). Biological identifications through DNA barcodes. *Proc. Roy. Soc. London, Ser. B, Biol. Sci.* 270: 313–321.

Hebert, P.D.N., Ratnasingham, S. & deWaard, J.R. (2003b). Barcoding animal life: Cytochrome c oxidase subunit 1 divergences among closely related species. *Proc. Roy. Soc. London, Ser. B, Biol. Sci.* 270: S96–S99.

Ho, J.K.I., Quek, R.F., Ramchunder, S.J., Memory, A., Theng, M.T.Y., Yeo, D.C.J. & Clews, E. (2018). Aquatic macroinvertebrate richness, abundance and distribution in the Nee Soon freshwater swamp forest, Singapore. *Gard. Bull. Singapore* 70 (Suppl. 1): 71–108.

Hollingsworth, P.M. (2008). DNA barcoding plants in biodiversity hot spots: Progress and outstanding questions. *Heredity* 101: 1–2.

Hollingsworth, P.M., Graham, S.W. & Little, D.P. (2011). Choosing and using a plant DNA barcode. *PLoS ONE* 6(5): e19254

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

Kwong, S., Srivathsan, A., Meier, R. (2012a). An update on DNA barcoding: low species coverage and numerous unidentified sequences. *Cladistics* 28: 639–644.

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

Lim, G.S., Balke, M. & Meier, R. (2011). Determining Species Boundaries in a World Full of Rarity: Singletons, Species Delimitation Methods. *Syst. Biol.* 61(1):165–169.

Lim, N.K.M., Tay, Y.C., Srivathsan, A., Tan, J.W.T., Kwik, J.T.B., Baloğlu, B., Meier, R. & Yeo, D.C.J. (2016). Next-generation freshwater bioassessment: eDNA metabarcoding with a conserved metazoan primer reveals species-rich and reservoir-specific communities. *Royal So. Open Sci.* 3: 160635.

Lim, W.H., Li, T.J. & Cai, Y. (2018). Terrestrial snails and slugs diversity in Nee Soon freshwater swamp forest, Singapore. *Gard. Bull. Singapore* 70 (Suppl. 1): 109–121.

Meier, R. (2008). DNA sequences in taxonomy - Opportunities and challenges. In: Wheeler, Q.D. (ed) *New Taxonomy*, pp. 95-127. London: CRC Press.

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. *Conserv. Biol.* 18: 478–488.

Meier, R., Shiyang, K., 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: 715–728.

Meier, R., Zhang, G.Y. & 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.

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.

Meyer, M. & Kircher, M. (2010). Illumina sequencing library preparation for highly multiplexed target capture and sequencing. *Cold Spring Harb. Protoc.* 2010: 5448.

Ng, T.H., Tan, S.K., Wong, W.H., Meier, R., Chan, S.Y., Tan, H.H. & Yeo, D.C.J. (2016). Molluscs for sale: assessment of freshwater gastropods and bivalves in the ornamental pet trade. *PLoS ONE* 11: e0161130.

Ødegaard, F. (2000). How many species of arthropods? Erwin’s estimate revised. *Biol. J. Linn. Soc.* 71: 583–597.

Pont, A.C. & Meier, R. (2002). *The Sepsidae (Diptera) of Europe*. *Fauna Entomologica Scandinavica*, vol. 37. Leiden: Brill.

Rohner, P.T., Ang, Y., Lei, Z. & 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: 555–563.

Srivathsan, A. & Meier, R. (2012). On the inappropriate use of Kimura-2-parameter (K2P) divergences in the DNA-barcoding literature. *Cladistics* 28: 190–194.

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

Srivathsan, A., Ang, A., Vogler, A.P. & Meier, R. (2016). Fecal metagenomics for the simultaneous assessment of diet, parasites, and population genetics of an understudied primate. *Front. Zool.* 13: 1–13.

Straub, S.C.K., Parks, M., Weitemier, K., Fishbein, M., Cronn, R.C. & Liston, A. (2012). Navigating the tip of the genomic iceberg: Next-generation sequencing for plant systematics. *Am. J. Bot.* 99: 349–364.

Tan, D.S.H., 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: 51–61.

Walter, D.E. & Winterton, S. (2007). Keys and the crisis in taxonomy: Extinction or reinvention? *Annu. Rev. Entomol.* 52: 193–208.

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. *Mol. Ecol. Resour.* https://doi.org/10.1111/1755-0998.12751

Wheeler, Q.D. & Meier, R. (2000). *Species Concepts and Phylogenetic Theory. A Debate.* New York: Columbia University Press.

Winston, J.E. (1999). *Describing Species: Practical Taxonomic Procedure for Biologists.* New York: Columbia University Press.

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

Yeo, D., Puniamoorthy, J., Ngiam R. W. J., Meier, R. (in press). Towards holomorphology in entomology: rapid and cost-effective larval-adult matching using NGS barcodes. *Sys. Ent.*

Zhang, J., Kobert, K., Flouri, T. & Stamatakis, A. (2014). PEAR: a fast and accurate Illumina Paired-End reAd mergeR. *Bioinformatics* 30(5): 614–620.