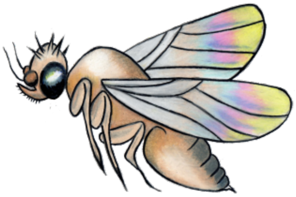
Stockholm
University

A multi-faceted approach to a "dark taxon"

The hyperdiverse and poorly known scuttle flies (Diptera: Phoridae)

Emily Hartop



Doctoral Thesis in Systematic Zoology at Stockholm University, Sweden 2021

A multi-faceted approach to a "dark taxon"

The hyperdiverse and poorly known scuttle flies (Diptera: Phoridae)

Emily Hartop

Academic dissertation for the Degree of Doctor of Philosophy in Systematic Zoology at Stockholm University to be publicly defended on Tuesday 8 June 2021 at 13.00 online via Zoom, public link is available at the department website.

Abstract

Most of the unknown animal biodiversity on earth is in groups of invertebrates that are hyperdiverse and abundant, yet poorly known ("dark taxa"). The study of these organisms requires a multi-faceted approach and methodologies designed to tackle large numbers of species and specimens. The scuttle flies (Diptera: Phoridae) are a classic example of a dark taxon and the focus of this thesis. Paper I is a molecular phylogeny of the phorid genus *Megaselia* based on one nuclear (28S rDNA) and three mitochondrial (ND1, COI and 16S) markers from 145 species of Nordic *Megaselia*. Molecular data was analysed with Bayesian analysis, maximum likelihood, and parsimony methods. Based on these results, and supporting morphological data, we divide *Megaselia* into 22 informal species groups, 20 of which fall into a monophyletic "core *Megaselia*". We discuss implications for the future circumscription of *Megaselia* and associated genera. Paper II presents a pipeline for rapid and cost-effective species discovery using the Oxford Nanopore mobile sequencing technology MinION. This paper reveals the presence of ca. 650 species of Phoridae from a single Malaise trap placed in Kibale National Park, Uganda. Based on our data, we estimate that the phorid fauna of the Afrotropical region could be as high as 100 000 species: this figure dwarfs previous diversity estimates. The implications for species discovery and description are discussed, and a new species (*Megaselia sepsioides* sp. nov.) is described. Paper III outlines a large-scale integrative approach to species discovery and delimitation in hyperdiverse groups, exemplified using a dataset of 18 000 phorid flies from Sweden. COI minibarcodes (313 bp) were obtained for all specimens and classified into putative species using different clustering methods (objective clustering, Poisson tree process, automatic barcode gap discovery and refined single linkage). No clustering method was accurate enough to use for species delimitation without confirmation from additional data. We found that the stability of a cluster to change across genetic-distance thresholds and the genetic variation within a cluster both accurately predict clusters where morphology is likely to be incongruent with barcode data. With molecular clustering integrated with morphological validation, we found that we could examine less than 5% of specimens and still delimit all species fully and accurately. Paper IV addresses questions about the scuttle fly fauna of Sweden with data from 32 000 scuttle flies from 37 sites and 4 time periods. We estimate that the total Swedish fauna contains 652-713 (based on Chao 1 or CNE estimates, respectively) species of scuttle flies, 1.5 times the 372 species currently documented from Sweden. Ordination techniques show that scuttle fly communities are organized in a gradient across Sweden, which is well correlated with plant hardiness zones defined by the Swedish Horticultural Society. Hierarchical modelling of species communities (HMSC) reveals that phorid community composition is largely determined by climatic and temporal variables, but much of the variance remains unexplained by the models we explored. Comparison of our phorid data with that of species more commonly utilised for biodiversity assessments revealed that phorids may allow more fine-scaled analysis as they may exist in smaller ranges, and that they additionally may give unique patterns of distribution that are unlike those seen in other taxa.

Keywords: *biodiversity, taxonomy, dark taxa.*

Stockholm 2021
http://urn.kb.se/resolve?urn=urn:nbn:se:su:diva-192276

ISBN 978-91-7911-472-5
ISBN 978-91-7911-473-2

Department of Zoology

Stockholm University, 106 91 Stockholm

A MULTI-FACETED APPROACH TO A "DARK TAXON"

Emily Hartop

Stockholm
University

A multi-faceted approach to a "dark taxon"

The hyperdiverse and poorly known scuttle flies (Diptera: Phoridae)

Emily Hartop

ISBN print 978-91-7911-472-5
ISBN PDF 978-91-7911-473-2

Scuttle fly cover art by Megan Zehendner.
Printed in Sweden by Universitetsservice US-AB, Stockholm 2021

Dedicated to my parents
whose unwavering
support started with my
childhood "dead bug
collection" – here are
some words about a
bigger dead bug
collection!

ACKNOWLEDGEMENTS

Thank you to my **family**: **Mom, Dad, Megan, Jason, Shade-Contraption, Uncle Chris, Aunt Juliet, Charlotte, Anthony, Z'Abigail, Darlene**, and to the memory of my **Nana** who saw the beginning of this journey but not the end – thank you for being the best worldwide support system a person could hope for.

To **Brian Brown**, thank you for getting me started on phorids and being a wonderful friend and mentor. You are one of the few true adventurers I know. Your passion and enthusiasm for natural history are much needed reminders of why I became an entomologist in the first place.

Thank you to **Henry Disney** for inspiring and encouraging me to persevere with Megaselia. Our many conversations over tea have helped me stay away from the physics department and come to terms with my life decision to study the impossible!

To the entire team at **Station Linné**, thank you for your support these past four years. A special thank you to **Dave Karlsson**, without whom so many things would not have been possible. To **Mathias and Catrin Jaschhof**, I will miss van campouts and Christmas Eve dinners…you have been such important friends and colleagues during my time in Sweden.

To my amazing friends in Vickleby: **Nina, Pelle, Phil, Cissi, Mira, Marcus, Sofia, Vilhelm,** and **Stella**. Thank you for being the very best community of people I have ever known. So many memories!

To **Fredrik Sjöberg**, it has been magical getting to know you. I will continue working on my Swedish just to muddle through your beautiful books.

Thank you to fellow phorid specialists **John Hash** and **Sibylle Häggqvist** for collaboration and consultation on all matters pertaining to the world's most awesome flies.

Thank you to all in the **Ronquist**, **Meier** and **Łukasik labs** for helping me navigate these last four years. A special thank you to **Allison Hsiang**, you helped me solve several tough problems when I was totally stuck!

Thank you to **Tomas Roslin**, **Mattias Forshage**, **Piotr Łukasik**, for advice, comments, and conversation.

Thank you to my co-authors **Darren Yeo**, **Mirkka Jones**, and **Sven Olof Ulefors**, working with you has allowed me to do so much more than I could have done alone, and I have learned so

much. Super thanks to **Leshon Lee**, I am so glad you are studying phorids, great things to come! A special thank you to my co-author and friend **Amrita Srivathsan**. You amaze me! Thank you for all your help. See you for a beer soon!

To my co-advisor, **Kjell Arne Johanson**, and my follow-up group members, **Chris Wheat** and **Kevin Holston** for the support and conversations these past four years.

To my advisor, **Fredrik Ronquist** – I am so grateful that you suggested this venture! Thank you for your time these last four years. I will continue to work on my R skills! Thank you to you and **Eva Ronquist** for hosting me on my visits to Stockholm. We still need to watch Holy Grail!

To **Rudolf Meier**, how on earth would I have ever made it through the *mar crudele* of a PhD without you? I cannot thank you enough. So much is ahead!

To **Cat**, you made being cooped up in an apartment writing for months on end bearable. Your constant fluffy presence really made a difference in morale.

Finally, to **Frank**. I could not have done this without your support. WE did it! I love you.

A MULTI-FACETED APPROACH TO A “DARK TAXON”: THE HYPERDIVERSE AND POORLY KNOWN SCUTTLE FLIES (DIPTERA: PHORIDAE)

CONTENTS

Abstract

Most of the unknown animal biodiversity on earth is in groups of invertebrates that are hyperdiverse and abundant, yet poorly known ("dark taxa"). The study of these organisms requires a multi-faceted approach and methodologies designed to tackle large numbers of species and specimens. The scuttle flies (Diptera: Phoridae) are a classic example of a dark taxon and the focus of this thesis. *Paper I* is a molecular phylogeny of the phorid genus *Megaselia* based on one nuclear (28S rDNA) and three mitochondrial (ND1, COI and 16S) markers from 145 species of Nordic *Megaselia*. Molecular data was analysed with Bayesian analysis, maximum likelihood, and parsimony methods. Based on these results, and supporting morphological data, we divide *Megaselia* into 22 informal species groups, 20 of which fall into a monophyletic "core *Megaselia*". We discuss implications for the future circumscription of *Megaselia* and associated genera. *Paper II* presents a pipeline for rapid and cost-effective species discovery using the Oxford Nanopore mobile sequencing technology MinION. This paper reveals the presence of ca. 650 species of Phoridae from a single Malaise trap placed in Kibale National Park, Uganda. Based on our data, we estimate that the *Megaselia* fauna of the Afrotropical region could be as high as 100 000 species: this figure dwarfs previous diversity estimates. The implications for species discovery and description are discussed, and a new species (*Megaselia sepsioides* sp. nov.) is described. *Paper III* outlines a large-scale integrative approach to species discovery and delimitation in hyperdiverse groups, exemplified using a dataset of 18 000 phorid flies from Sweden. COI minibarcodes (313 bp) were obtained for all specimens and classified into putative species using different clustering methods (objective clustering, Poisson tree process, automatic barcode gap discovery and refined single linkage). No clustering method was accurate enough to use for species delimitation without confirmation from additional data. We found that the stability of a cluster to change across genetic-distance thresholds and the genetic variation within a cluster both accurately predict clusters where morphology is likely to be incongruent with barcode data. With molecular clustering integrated with morphological validation, we found that we could examine less than 5% of specimens and still delimit all species fully and accurately. *Paper IV* addresses questions about the scuttle fly fauna of Sweden with data from 32 000 scuttle flies from 37 sites and 4 time periods. We estimate that the total Swedish fauna contains 652-713 (based on Chao 1 or CNE estimates, respectively) species of scuttle flies, 1.5 times the 372 species currently documented from Sweden. Ordination techniques show that scuttle fly communities are organized in a gradient across Sweden, which is well correlated with plant hardiness zones defined by the Swedish Horticultural Society. Hierarchical modelling of species communities (HMSC) reveals that phorid community composition is largely determined by climatic and temporal variables, but much of the variance remains unexplained by the models we explored. Comparison of our phorid data with that of species more commonly utilised for biodiversity assessments revealed that phorids may allow more fine-scaled analysis as they may exist in smaller ranges, and that they additionally may give unique patterns of distribution that are unlike those seen in other taxa.

SAMMANFATTNING PÅ SVENSKA

Den biologiska mångfalden på vår planet består till stor del av ryggradslösa djur, bland vilka ett antal vanligt förekommande och synnerligen artrika grupper tillhör de allra sämst kända. Med ett fackuttryck kallas de senare ”dark taxa”. Att studera sådana grupper kräver helt andra metoder än de som vanligen används inom den taxonomiska forskningen. Man behöver hantera både det stora antalet arter och den stora mängden exemplar i materialet man samlar in. Puckelflugorna (Diptera: Phoridae), som denna avhandling fokuserar på, är ett typiskt exempel på ett ”dark taxon”. Puckelflugor är en grupp små flugor som är talrika i de flesta miljöer. Deras biologi är mycket varierande: de kan vara parasiter på olika insekter och andra leddjur, men också nedbrytare eller svampätare. Puckelflugorna är en av de artrikaste och vanligaste insektsfamiljerna i Sverige, liksom i många anda delar av världen.

I den första uppsatsen i avhandlingen (*uppsats I*) analyserar vi släktskapsförhållandena i det mest artrika släktet inom familjen, släktet *Megaselia*. Släktet omfattar mer än 1600 beskrivna arter. Det inbegriper de flesta arter inom familjen och hör till ett av de mest artrika släktena i hela djurriket. Vår analys är baserad på en nukleär markör (28S rDNA; en gen från DNA i cellkärnan) och tre mitokondriella markörer (ND1, COI och 16S), som vi sekvenserat från 145 nordiska *Megaselia*-arter och några närbesläktade referensgrupper, så kallade utgrupper. Data analyserades med tre olika typer av metoder: Bayesiansk analys, maximum likelihood och parsimoni. Baserat på resultaten, och med stöd av morfologiska data, delar vi in släktet *Megaselia* i 22 informella artgrupper, av vilka 20 bildar en monofyletisk kärna (”core *Megaselia*”). Vi resonerar i uppsatsen kring vad resultaten skulle kunna betyda för kommande systematiska bearbetningar av *Megaselia* och närstående släkten. Till exempel skulle den monofyletiska kärnan vi identifierat kunna ligga till grund för en mer precis avgränsning av släktet. Även om vår analys behöver kompletteras med betydligt fler arter och fler genetiska markörer, så utgör den ett viktigt steg mot en bättre förståelse av de naturliga grupperingarna inom det enorma släktet *Megaselia*.

I den andra uppsatsen (*papper II*) beskriver vi en snabb och kostnadseffektiv metod för att hitta nya, obeskrivna arter med hjälp av Oxford Nanopores mobila sekvenseringsteknik MinION. I artikeln visar vi att ett material insamlat i en enda Malaisefälla i Kibale National Park i Uganda under åtta veckor innehåller cirka 650 arter av puckelflugor. Baserat på detta uppskattar vi att *Megaselia*-faunan i den afrotropiska regionen kan omfatta så många som 100 000 arter —en siffra som vida överstiger tidigare uppskattningar. Även betydelsen av nyupptäckta arter och beskrivningar av dessa diskuteras och en ny art (*Megaselia sepsioides* sp. nov.) beskrivs.

I den tredje uppsatsen (*papper III*) utvecklar vi en storskalig, integrerande metod för att snabbt hitta och avgränsa arterna i stora material av megadiversa grupper (”dark taxa”). Metoden bygger på snabb och billig individbaserad sekvensering av små genetiska markörer från ett stort antal exemplar (som i *papper II*), kombinerat med detaljerade studier av några få utvalda exemplar för att snabbt kunna fastställa artgränserna. Vi prövar olika varianter av metoden på data från 18 000 puckelflugor insamlade i Sverige. Vi tog först fram data för korta sekvenser av COI (313 bp; ofta kallade ”ministreckkoder” eller ”minibarcodes” på engelska för att de är så användbara för att skilja arter) från alla exemplar. Enskilda sekvensvarianter (”haplotyper”) grupperades sedan i presumtiva arter med hjälp av olika klustermetoder (”objective clustering”, ”Poisson tree process”, ”automatic barcode gap discovery” och ”refined single linkage”), och vi testade om grupperna överensstämde med morfologiskt identifierade arter. Ingen enskild metod var tillräckligt utslagsgivande för att ensamt kunna användas för artavgränsning utan kompletterande analyser av ytterligare data. Vi fann att både stor genetisk variationen inom ett kluster och hur robust klustret är för ändringar av tröskelvärdet i klusteralgoritmen kunde förutsäga de kluster som inte överensstämmer med morfologin med god precision. Genom att kombinera preliminär klustring av streckkoder med morfologisk validering fann vi att vi behövde undersöka färre än 5% av exemplaren i materialet för att kunna både hitta och avgränsa alla arter korrekt.

I den sista uppsatsen i avhandlingen (*papper IV*) analyserar vi data från 32 000 puckelflugor, insamlade från 37 platser olika platser i Sverige och i 4 olika tidsperioder, för att öka kunskapen om faunans storlek och sammansättning i Sverige. Analyserna av detta material indikerar att vårt land hyser mellan 652 och 713 arter puckelflugor (baserat på Chao 1 respektive CNE-estimat). Detta är mer än 1,5 gånger så många som de 372 arter som rapporterats från Sverige tidigare. Statistiska analyser visar att olika arter puckelflugor är grupperade i en gradient, som korrelerar väl med de odlingszoner som Riksförbundet Svensk Trädgård har definierat. Hierarkisk modellering av artsamhällen (HMSC) ger vid handen att artsammansättningen puckelflugor till stor del bestäms av klimat- och tidsvariabler, men mycket av variationen förblir oförklarad av de modeller vi analyserade. Våra analyser visar att puckelflugorna ger en mer högupplöst bild av den svenska naturen än de grupper som oftast används i den här typen av analyser. Detta eftersom de är betydligt mer artrika, och varje arts förekomst är mer begränsad rumsligt och tidsmässigt. Våra analyser antyder också att puckelflugearternas utbredningsmönster skiljer sig på viktiga sätt från dem vi ser hos mer välstuderade grupper.

Sammanfattningsvis bidrar vi i avhandlingen både till att lägga grunden till mer detaljerade systematiska studier av puckelflugorna, samtidigt som vi tar viktiga steg i utvecklingen av en helt ny metodik som kan få stor betydelse för att accelerera kartläggningen av mångfalden hos såväl puckelflugor som många andra dåligt kända och samtidigt artrika grupper. Den sista uppsatsen visar också att studier av den här typen av grupper har mycket att ge när det gäller förståelsen av hur våra ekosystem är sammansatta och hur de fungerar, kunskap som är kritisk för att vi ska kunna bevara dem för framtiden.

AUTHOR'S CONTRIBUTIONS

The thesis is based on the following articles, which are referred to in the text by their Roman numerals:

I **Hartop E.***, Häggqvist S.*, Ulefors S.-O., Ronquist F. (2021) Scuttling towards monophyly: phylogeny of the mega-diverse genus *Megaselia* (Diptera: Phoridae). *Systematic Entomology*. 46.1:71-82. https://doi.org/10.1111/syen.12448

II Srivathsan A., **Hartop E.A.**, Puniamoorthy J., Lee W.T., Kutty S.N., Kurina O., Meier R. (2019) Rapid, large-scale species discovery in hyperdiverse taxa using 1D MinION sequencing. *BMC Biology*. 17:96. https://doi.org/10.1186/s12915-019-0706-9

III **Hartop, E.**, Srivathsan, A., Ronquist, F., Meier, R. (2021) Large-scale Integrative Taxonomy (LIT): resolving the data conundrum for dark taxa. bioRxiv. 2021.04.13.439467. https://doi.org/10.1101/2021.04.13.439467

VI **Hartop, E.**, Lee, L., Srivathsan, A., Jones, M., Yeo, D., Meier, R. (2021) A first dive into the terrestrial deep-sea trenches of Sweden: species richness, spatiotemporal distribution, and community composition of a dark taxon (Diptera: Phoridae). *Manuscript*.

*Joint first authorship

Candidate contributions to thesis articles*

	I	II	III	IV
Conceived the study	3	3	1	1
Designed the study	1	3	1	1
Collected the data	3	1	1	1
Analysed the data	1	2	1	1
Manuscript preparation	1	1	1	1

*** Contribution Explanation**

3 Minor: contributed in some way, but contribution was limited.

2 Significant: provided a significant contribution to the work.

1 Substantial: took the lead role and performed the majority of the work.

INTRODUCTION

Små, svarta, svårbeständа, charmlösa puckelflugor – Fredrik Sjöberg, Oslo, 2019

Much of the multicellular biodiversity on earth belongs to large groups of small invertebrates that are both diverse and abundant. These groups are often referred to as "dark taxa", a term originally coined to refer to the growing numbers of sequences in public databases that were not linked to species names (Page 2011, 2015). In current usage, "dark taxa" more specifically refers to species rich groups for which a large proportion of species are undescribed. Studying these groups presents myriad challenges, and we often lack information that is considered basic knowledge for less diverse groups. For example, we often do not know, or have limited knowledge of, the numbers, evolution, composition, delimitation, and distribution of species in these hyperdiverse clades. This is due to the compounding factors of abundance and diversity: when both specimen and species numbers are high, the processes of discovering and delimiting species become extremely challenging. These are further complicated by often challenging morphology and small size. It is no surprise that these groups are often largely made up of undescribed species, and we know little about even the (comparatively) small numbers of those that have been described.

It is important to prioritise the study of these groups, as they make up most of the terrestrial species diversity and large amounts of biomass. They are likely crucial to the functioning of many ecological processes but are so understudied that we cannot begin to assess their impact. As anthropogenic change alters our planet at an alarming rate, we must prioritise understanding those groups that are quantitatively indicated as critical as the future wellbeing of our society is dependent on the natural resources and services they provide (Swiss Re Institute 2020).

Studying hyperdiverse groups requires a different approach than studying most taxa. The development and implementation of tools and workflows that are not required for more "normal" groups is critical. Methods must be able to organise and simplify the large numbers of species and specimens in dark taxa to make the inaccessible accessible. Here, we present a series of papers that take a multi-faceted approach to a dark taxon in Sweden. This thesis represents a progression of knowledge, from an evolutionary backbone in *paper I*, to new and innovative methodologies for handling dark taxa in *papers II* and *III*, to a first look at the richness, diversity, and spatiotemporal distributions of a dark taxon in Sweden in *paper IV*.

Sweden has one of the best studied faunas (and floras) on earth. Sweden was home to Linnaeus, who pioneered the centuries of focused taxonomic study that would follow, aimed at discovering and describing biodiversity. Despite progress that was made, some groups of organisms remained poorly known (dark taxa). In recent years, the Swedish Taxonomy Initiative (STI) has focused efforts on such groups, facilitating advancements of their study by funding projects like some of the work in this thesis.

Here, we use scuttle flies (Diptera: Phoridae) as an example of a dark taxon. This family is comprised of over 4 000 described species but is dominated by the genus *Megaselia*, especially in temperate regions. First described over 160 years ago, *Megaselia* contains approximately 1 700 formalised species. Despite the description of so many species, we have not achieved anything close to a comprehensive understanding of the genus, and *Megaselia* is estimated to contain tens, if not hundreds, of thousands of species (Brown et al. 2018; Srivathsan et al. 2019). Unfortunately, we not only lack an agreed upon estimate of species numbers, we also lack an understanding of the evolutionary history and boundaries of this group, and we have no confident estimate as to the true numbers of species or their distributions across time and space.

Megaselia is one of the most biologically diverse genera in the animal kingdom (Marshall 2012) and, as with many dark taxa, it is a group that is certainly important to the function of ecosystems. The known species of *Megaselia* are well documented as having multitudinous life histories across a wide range of habitats (Disney 1979, 1990, 1994). Such a combination – of species richness and abundance coupled with biological diversity – makes *Megaselia* a potential goldmine of ecological data. To make *Megaselia* (and the family Phoridae more generally) accessible for biodiversity assessments, however, we must be able to efficiently create and analyse large datasets on these groups. This requires a much better understanding of the genus and family than we have at present.

The approach to scuttle flies herein is an attempt to answer a series of basic questions about this family of flies, and to develop methods designed to facilitate this. These methods are not specific to scuttle flies; they are designed to address challenges faced in all hyperdiverse/dark taxa. The problems are, chiefly, the diversity of species and the abundance of specimens. Each paper addresses a particular question about scuttle flies (or the genus *Megaselia,* in particular), and papers II-IV additionally use new and innovative methods developed and optimised for use on hyperdiverse taxa.

In *paper I*, we address the need for an evolutionary framework for *Megaselia*. The subfamily Metopininae, in which *Megaselia* resides, is poorly understood itself. Previous morphological analyses of the subfamily have suggested that *Megaselia* is paraphyletic with respect to several other genera, but a comprehensive analysis of the subfamily has yet to be completed (Disney, 1989).

Some classifications of *Megaselia* have been attempted in the past based on morphological characters. A separation of the genus into two subgenera (*Megaselia* and *Aphiochaeta*) by Enderlein (1924) was followed by further breakdowns into divisions and rows (Schmitz 1953, 1955, 1956, 1957, 1958). Unfortunately, Schmitz died before completing his work on *Megaselia* and even with continued efforts by others (Schmitz and Beyer 1965a, 1965b; Schmitz and Delage 1974, 1981), the study of some divisions of *Megaselia* was never completed. The groups proposed by Enderlein and Schmitz have long been used by *Megaselia* specialists as practical, but unnatural, divisions (Disney 1994). Despite attempts at categorising *Megaselia*, it has remained a "dustbin" for species that cannot be put elsewhere, based on a loose characterisation that is made up largely of negative characters. An evolutionary framework has long been needed to define and restrict *Megaselia* and, further, to begin to organise the species within into manageable subunits.

In *paper I* we present the first attempt at a large molecular phylogeny for *Megaselia* that has ever been completed. We used four molecular markers sequenced from 175 Nordic specimens that represented 145 species of *Megaselia*. The phylogeny reveals 22 well-supported species groups within the current boundaries of the genus, including 20 that fall within a moderately well supported monophyletic "core". The two species groups that fall outside of the core of *Megaselia* are herein diagnosed, this should help to focus future work on circumscribing *Megaselia* and related genera.

Prior to this study, the lack of an evolutionary framework for *Megaselia* has discouraged many from taking up the study of the genus. With no way to focus on smaller subunits or monophyletic clades, most specialists have been unwilling to tackle *Megaselia* in its entirety (wise). With *paper I* we hope to facilitate future studies on *Megaselia*, but we still need methods to efficiently discover and delimit species within this framework. *Paper II* therefore focuses on a method to discover new species from large molecular datasets.

In *paper II*, we address the problems of discovering species in hyperdiverse/dark taxa through an efficient and cost-effective molecular pipeline. Traditionally, species discovery is a slow process. In species rich and specimen-abundant groups like *Megaselia* it can be extremely slow. To find new or rare species within large insect samples that contain thousands, or even tens-of-thousands of insects, the samples must be sorted. Preliminary sorting to higher taxonomic levels can be done by workers with basic training (students or technicians), but once species-level sorting is required, the work must be completed by experts. This often involves time consuming techniques like painstaking dissections, specimen mountings (e.g., slides), and detailed drawings. In groups like *Megaselia*, these processes can be formidably time consuming, making even the identification of a species as potentially new a drawn-out process.

Thankfully, in recent years workflows that utilise rapid and cost-effective NGS barcoding to pre-sort large samples into putative species have been developed and optimised (Srivathsan et al. 2018; Wang et al. 2018a; Srivathsan et al. 2019; Yeo et al. 2020; Srivathsan et al. 2021). With the development of this type of "reverse workflow", barcoding can be completed on thousands of specimens in short periods of time. Molecular barcodes can then be clustered using distance- or phylogeny-based algorithms, making quick work out of sorting large numbers of specimens into putative species. After initial sorting, experts can concentrate their efforts on species that are potentially new or interesting. In *paper II*, we demonstrate such a pipeline on 7 059 phorid specimens from Kibale National Park, Uganda. Specimens were collected over eight weeks from a single Malaise trap, and after sequencing were revealed to contain >650 species. This is more species of phorid than have previously been recorded in the entire Afrotropical region. We estimate that continued processing from this single trap will reveal this site to be home to upwards of 1 000 species of phorids and that the Afrotropical region may contain over 100 000 species of *Megaselia*, dwarfing previous diversity estimates. To demonstrate how this pipeline streamlines well-informed, efficient, and inexpensive species discovery, we describe a new species (*Megaselia sepsioides* sp. nov.).

The study for *paper II* was completed using a portable sequencer, the Oxford Nanopore MinION. In the study herein, we optimised this technology to a capacity of 3 500 barcodes per flowcell. Since the publication of *paper II* in 2019, advancements have been made to MinION technology that have allowed work to continue optimising this platform for large-scale barcoding. Now, MinION is truly the ideal tool for low cost, high throughput sequencing for species discovery – up to 10 000 barcodes can be sequenced on a single flowcell (Srivathsan et al. 2021). MinION is finally a perfect tool to bring barcoding to everyone – the processes are streamlined, accessible bioinformatics tools have been developed and made available, and the costs are now low enough to make this technology truly affordable (Srivathsan et al. 2021).

In *paper III* we took inspiration from the discovery of *Megaselia sepsioides* in *paper II* to apply large-scale barcoding to ca. 18 000 scuttle flies from across Sweden. Our goal was to develop a systematic and objective approach to large molecular datasets that would streamline the species discovery and delimitation processes while integrating multiple data sources. Integrative taxonomy has long been the gold standard in taxonomy, but systematic, efficient and formalised approaches are still underdeveloped (but see Puillandre et al. 2012b; Kekkonen and Hebert 2014). Instead, integrative taxonomy had most often relied on time consuming morphological sorting prior to the barcoding of representative specimens (Butcher et al. 2012; Riedel et al. 2013a; Lücking et al. 2016) or alternatively, some authors have equated molecular units to species without standardised methods of validation (Hebert et al. 2016; Sharkey et al. 2021). Either method is undesirable. Morphological sorting and representative sequencing is time consuming and is likely to miss some species, while delimiting based on barcodes without set validation procedures is not advised even by those who develop such methods (Puillandre et al. 2012a; Ratnasingham and Hebert 2013; Zhang et al. 2013).

Paper III therefore presents a solution to this integrative taxonomy conundrum by developing LIT: Large-scale Integrative Taxonomy. LIT is an approach that utilises a first data source that is cheap and easy to obtain to pre-sort large numbers of specimens into putative species. It then uses a second, more expensive, data source to validate species units. In our study, we used NGS minibarcodes as the first source and morphology as the second to develop LIT using ca. 18 000 barcodes from Swedish phorids. We first obtained NGS barcodes for all specimens and then used different molecular species delimitation methods and thresholds to create barcode clusters. In a first stage, we randomly picked 100 clusters to test whether we could identify cluster-specific traits (e.g., number of haplotypes, maximum pairwise distance) that would be able to predict whether a molecular cluster is likely to be incongruent with morphology. Once we identified these predictors, we tested their effectiveness when applied to the remaining barcode clusters. We then developed explicit rules for picking specimens from within each cluster that should be studied for validating the preliminary species hypotheses that were based on barcode clusters. We compared different clustering methods and concluded that none was sufficient on its own to circumscribe all species correctly. We finally formalise our system in an algorithm to demonstrate that it is systematic, objective, and effective.

In *paper IV* we wanted to use a large barcode dataset to address some fundamental questions about a dark taxon. As we witness dramatic changes to our planet in the Anthropocene, it becomes increasingly critical to monitor our biodiversity with quantitative priorities. This means addressing ecological questions with data from organisms that represent large portions of the species diversity and biomass, in additional to representing a broad spectrum of life history strategies. Unfortunately, many of the groups that would offer such abundant data are dark taxa that are poorly known and difficult to study.

For this reason, studies often use better known, more charismatic taxa for assessments. Studies based on organisms like birds and butterflies may or may not present accurate representations of biodiversity more broadly. To assess how different phorid results might be, we utilised the same barcode set as in *paper III* and sequenced a further ca. 14 000 barcodes for a total of ca. 32 000 barcodes. The additional barcodes were from the same locations as in *paper II* but represented four time-periods (late spring, midsummer, late summer and offseason). With this expanded dataset, we assessed the richness and distribution of phorids across space and time in Sweden. We additionally wanted to assess if phorids, as a representative dark taxon, showed patterns that would offer a unique perspective on biodiversity when compared to organisms that are more commonly utilised in biodiversity assessments. Finally, we wanted to test whether we needed to use species units for studying phorid distributions, or whether genetic diversity (haplotypes) would yield comparable results.

With all papers in combination, we here present a multi-faceted approach to studying a dark taxon. With the evolutionary framework presented in *paper I* combined with the richness and distribution data obtained in *paper IV*, we have a much better understanding of the phorid fauna of Sweden. We have additionally developed and optimised large-scale molecular workflows as well as a systematic and objective approach to morphological validation of the large-scale molecular datasets obtained with such workflows (*papers I* and *II*). All the tools and information in these four papers combined not only greatly advance our knowledge of this enigmatic and fascinating group, but they present ways forward to move beyond these studies to further taxonomic work (descriptions) and more in-depth ecological analyses.

Paper I

Materials and Methods

We sequenced 145 species of Nordic *Megaselia* that represent a broad sampling across *Megaselia* groups historically recognized by Schmitz (1956) and morphological working groups defined by Sven Olof Ulefors (SOU). SOU developed these working groups through the examination of ca. 35 000 specimens from the Swedish Malaise Trap Project, an inventory of the Swedish insect fauna conducted at 55 localities across Sweden between 2003-2006 (Karlsson et al. 2020). Specimens from nine other genera were selected as outgroups. Five were selected from within the Metopininae and four from the Phorinae.

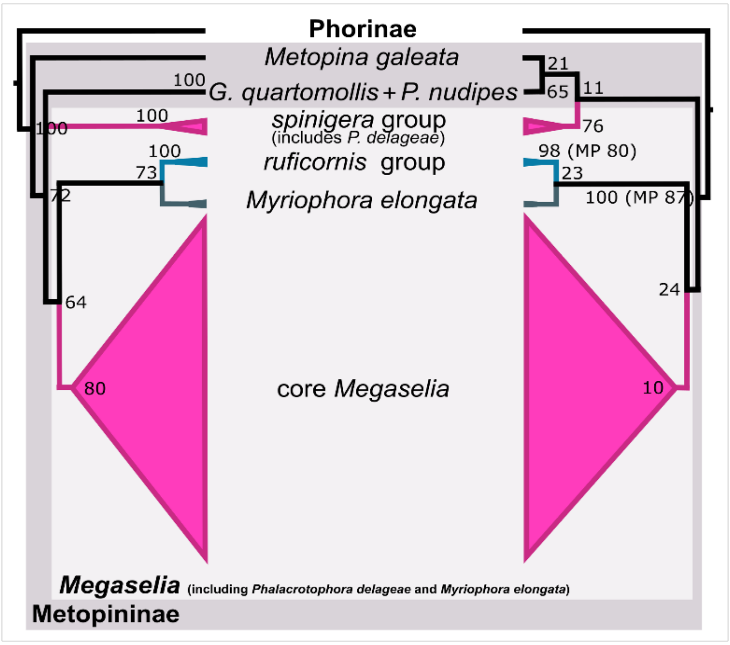


Figure 1. Condensed consensus tree showing generic relationships and basal clades based on Bayesian (left) and maximum likelihood (right) analyses. Maximum parsimony values are indicated on the ML tree in parentheses.

Our results are based on a combined analysis of four molecular markers: one nuclear – the D2 variable expansion region of nuclear 28S rDNA (28S, 527 bp) – and three mitochondrial – the barcode region of cytochrome oxidase I (COI, 658 bp), NADH1 dehydrogenase (ND1, 378 bp) and 543 bp of 16 S ribosomal RNA (16S), for a total of 2.1 kb of sequence data. Phylogenetic analyses were conducted using Bayesian analysis with MrBayes 3.2.6 (Ronquist et al. 2012), maximum likelihood analysis with RAxML (Stamatakis 2014), and maximum parsimony analysis with TNT (Goloboff and Catalano 2016). Rogue taxon analysis was conducted using RogueNaRok (Aberer et al. 2013).

Results

Our results confirm the monophyly of the Metopininae across all methods of analysis (Fig. 1). Furthermore, they reveal four major clades within *Megaselia*: the *spinigera* and *ruficornis* group of *Megaselia*, *Myriophora elongata*, and "core *Megaselia*" (Fig. 1). The *ruficornis* group is strongly supported across methods and is additionally morphologically diagnosed by the presence of differentiated intra-alar setae (Fig. 2a). The *spinigera* group is supported by both Bayesian and ML analyses and morphologically diagnosed by differentiated setae on the posterior margin of the scutum (Fig. 2b). Core *Megaselia* contains 20 species groups (Fig. 3) and is diagnosed morphologically by reduced setation on the scutum, as seen in generic type species *Megaselia costalis* (Fig. 2c). The ability to diagnose species within the "core" may, in future, mean that *Megaselia* will be restricted to this clade, making the genus monophyletic at last.

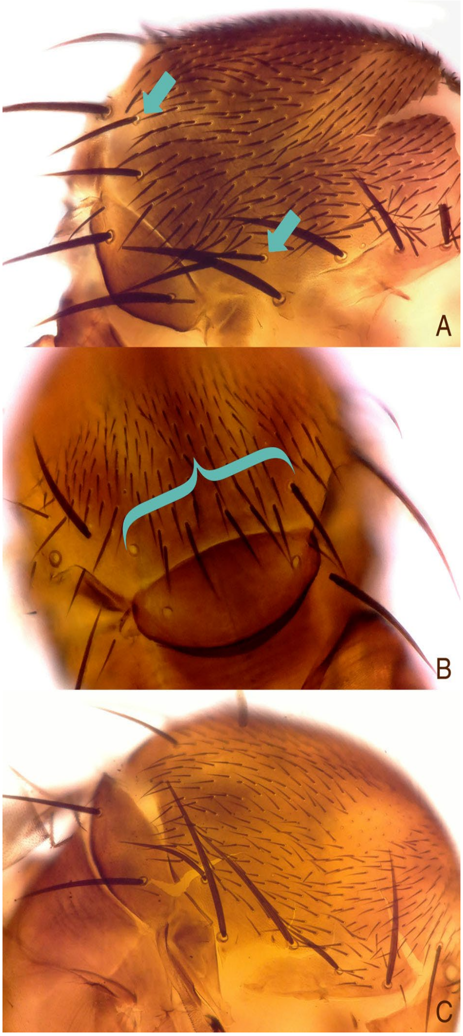


Figure 2. (a) Strongly differentiated intra-alar setae (marked with arrows) as found in the *ruficornis* group and *Myriophora*; (b) row of differentiated setae on posterior margin of scutum, between posterocentral setae, as in the *spinigera* group; (c) generic type species *Megaselia costalis* with the reduced setation proposed as a potentially diagnostic feature of core *Megaselia*.

Core *Megaselia* contains most of the species diversity sampled and is divided into 20 strongly supported species groups and eight species placed individually on the tree (Fig. 3). Relationships between these groups were poorly resolved in all analyses, and rogue taxon analysis did not make any major improvements to the phylogeny. We give potentially useful morphological characters for all groups, although most clades cannot be reliably diagnosed by morphology at this time.

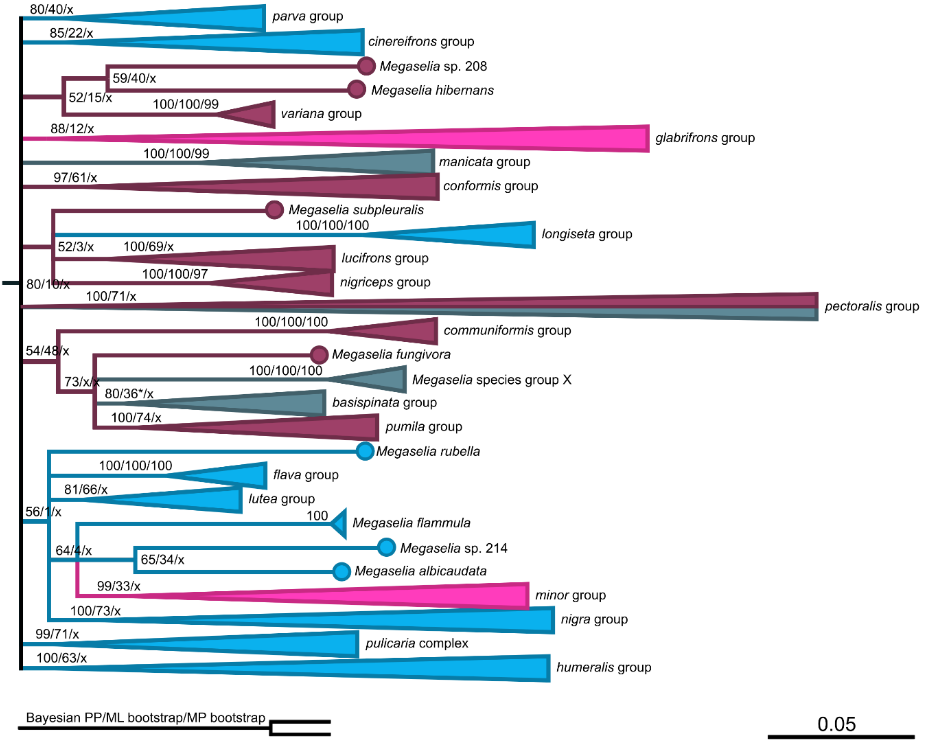


Figure 3. Consensus tree for core *Megaselia* showing 20 well supported species groups. Colours of the clades correspond to setosity of the anepisternum. Blue: anepisternum bare, pink: mixed group with some taxa with anepisternum bare and some setose, grey: anepisternum setose + differentiated bristle(s).

PAPER II

Materials and Methods

8,669 phorid flies were collected from a single site in Kibale National Park, Uganda (Kurina 2012) and full-length COI barcodes (658 bp) were sequenced using tagged amplicons building on the reverse workflow of Wang et al. (2018a). Sequencing was done with Oxford Nanopore's MinION sequencer using 1D library preparation. Two experiments were conducted. The first experiment ran for 48 hours and weak products were then identified, re-pooled and run on a new flowcell. The second experiment, endeavouring to lower costs and improve success rates, ran for 24 hours, then the flowcell was flushed and weak products were re-pooled and run again on the same flowcell.

After base-calling and demultiplexing, we aligned reads using MAFFT v7 (Katoh and Standley 2013). This was performed on a random subset of 100 reads from each demultiplexed file, a majority rule consensus was obtained, and all barcodes containing more than 1% ambiguous nucleotides were discarded. MAFFT barcodes were then corrected with RACON (Vaser et al. 2017) and both MAFFT and RACON barcodes were further corrected with an amino-acid correction pipeline (Srivathsan et al. 2018). We then consolidated these two sets of barcodes. A contamination check was performed using BLAST and 6 251 specimens were also sequenced on the Illumina platform for comparison. Finally, a morphological check for congruence was performed on 100 randomly selected mOTUS.

Methods are diagrammed in Figure 4.

Results

We obtained 7 273 preliminary barcodes from the 8,699 extracted specimens. This total was the product of two MinION runs that included re-pooling and re-sequencing of weak products (Fig. 5).

After MAFFT and RACON correction steps, 7 221 barcodes remained. Amino acid (AA) correction further reduced the barcode set to 7 178 AA-corrected MAFFT barcodes and 7 194 AA-corrected RACON barcodes. The two sets were then consolidated into 7 155 barcodes that, after contamination checks of the negative PCR controls and of the barcode set using BLAST, we obtained a final yield of 7 059 barcodes.

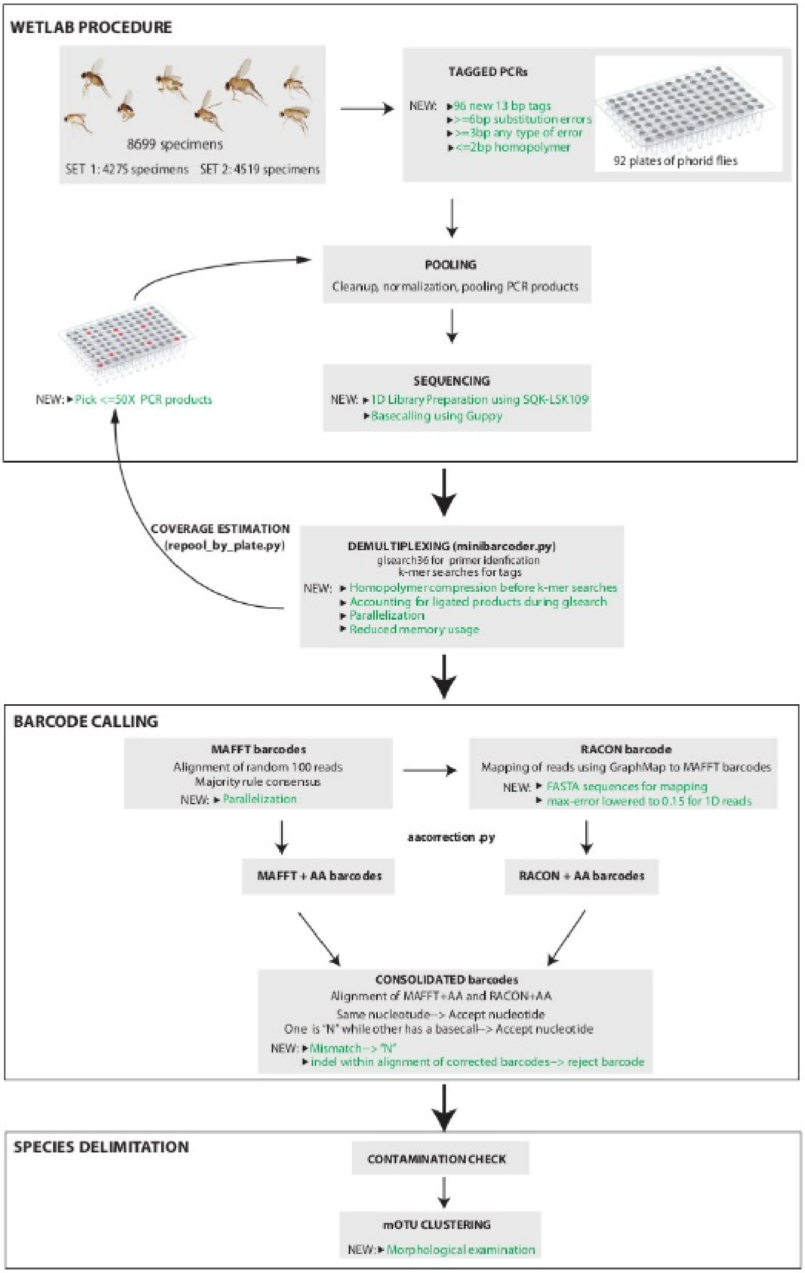

WETLAB PROCEDURE
8699 specimens
SET 1: 4275 specimens SET 2: 4519 specimens
TAGGED PCRs
NEW: ▸96 new 13 bp tags
▸>=6bp substitution errors
▸>=3bp any type of error
▸<=2bp homopolymer
92 plates of phorid flies
POOLING
Cleanup, normalization, pooling PCR products
SEQUENCING
NEW: ▸1D Library Preparation using SQK-LSK109
▸Basecalling using Guppy
NEW: ▸Pick <=50X PCR products
COVERAGE ESTIMATION
(repool_by_plate.py)
DEMULTIPLEXING (minibarcoder.py)
glsearch36 for primer idenfication
k-mer searches for tags
NEW: ▸ Homopolymer compression before k-mer searches
▸ Accounting for ligated products during glsearch
▸ Parallelization
▸ Reduced memory usage
BARCODE CALLING
MAFFT barcodes
Alignment of random 100 reads
Majority rule consensus
NEW: ▸ Parallelization
RACON barcode
Mapping of reads using GraphMap to MAFFT barcodes
NEW: ▸ FASTA sequences for mapping
▸ max-error lowered to 0.15 for 1D reads
aacorrection.py
MAFFT + AA barcodes
RACON + AA barcodes
CONSOLIDATED barcodes
Alignment of MAFFT+AA and RACON+AA
Same nucleotude--> Accept nucleotide
One is "N" while other has a basecall--> Accept nucleotide
NEW: ▸Mismatch--> "N"
▸ indel within alignment of corrected barcodes--> reject barcode
SPECIES DELIMITATION
CONTAMINATION CHECK
mOTU CLUSTERING
NEW: ▸ Morphological examination

Figure 4. MinION pipeline flowchart.

We obtained 6 251 mini barcodes (313 bp) via Illumina from the 7 059 specimens represented in the final dataset. Our consolidated MinION barcodes were found to be 99.99% accurate when compared to the Illumina results (accepting those as the true barcodes). Comparison of the mOTUs generated from the two platforms had a match ratio of 0.951 when compared at the same percentage, but all clusters were congruent when compared at 1.9-3.7%.

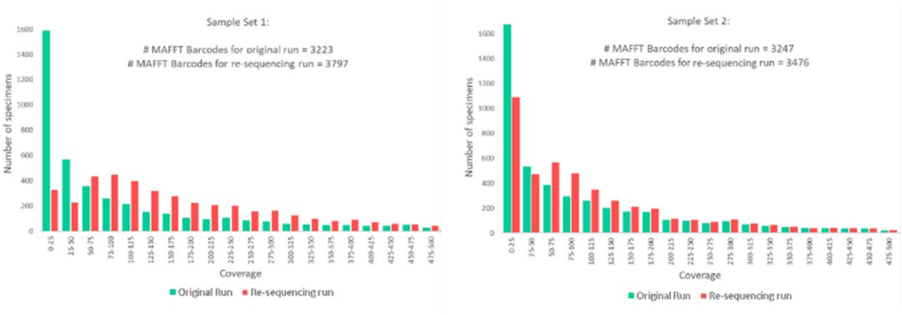


Figure 5. Effect of re-pooling on barcode coverage.

We clustered barcodes at a priori thresholds between 2 and 4% minimum interspecific distance (Cbjective Clustering, part of TaxonDNA in Meier et al. 2006)) and obtained between 613 and 705 mOTUs (2%: 705, 3%: 663, 4%: 613). An alternative species delimitation based on Poisson Tree Processes (Zhang et al. 2013) yielded a higher estimate of 747 mOTUs. We calculated the Chao 1 species richness estimate based on 3% mOTUs, this suggested the diversity exceeds 1 000 species at the single site sampled (Fig. 6).

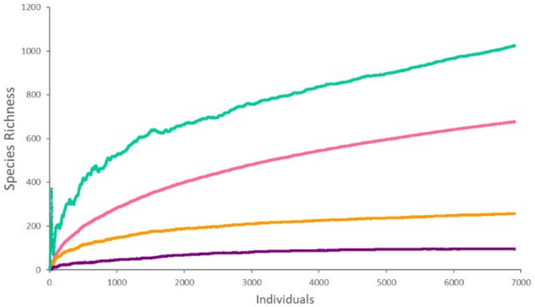


Figure 6. Species richness estimation. Green: Chao1 Mean, Pink: S (Mean), Orange: Singleton Mean, Purple: Doubleton Mean

We conducted a morphological validation step, checking all specimens in 100 randomly selected mOTUS. We found molecular-morphological congruence for 93% of mOTUs and >99% of specimens. 90% of the putative species were identified as belonging to the hyperdiverse genus *Megaselia*.

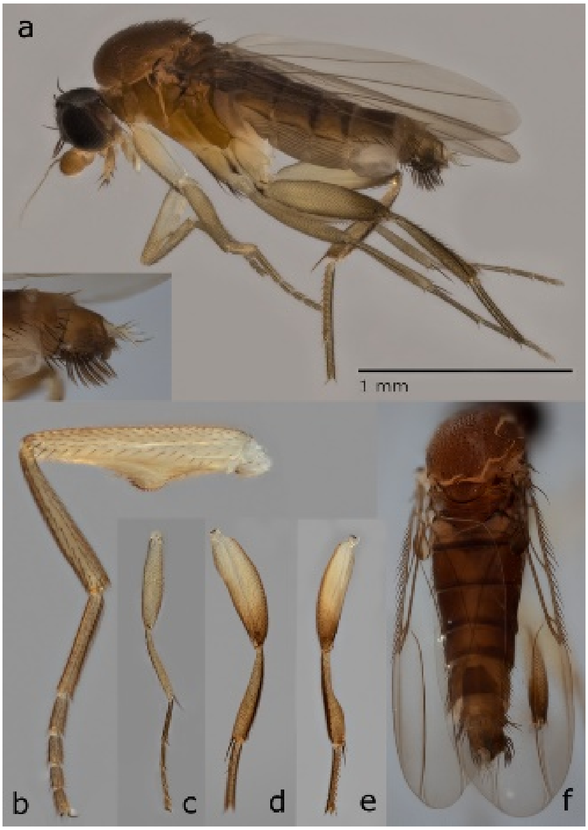


Figure 7. Lateral habitus and diagnostic features of *Megaselia sepsioides* sp. nov.

During the morphological validation process, we identified a distinctive new species that we described as *Megaselia sepsioides* (Fig. 7). This was a practical demonstration of how this workflow can guide well-informed species descriptions that include documentation of both molecular and morphological variation (Fig. 8).

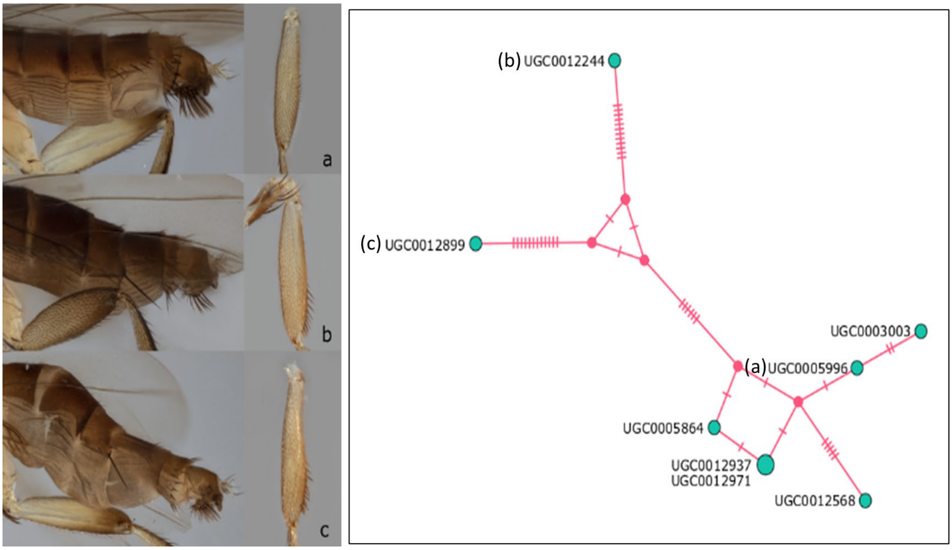


Figure 8. Documentation of morphological (left) and molecular (right) variation in *Megaselia sepsioides* sp. nov.

Paper III

Materials and Methods

Samples were collected at thirty-six sites across Sweden as part of the Swedish Insect Inventory Project (Fig. 9a) (Karlsson et al. 2020). Phorid flies were sorted from a summer sample selected from each site and sequenced to obtain 313 bp minibarcodes. Barcodes were clustered at 3% minimum interspecific distance threshold (Objective clustering (OC) in Meier et al. 2006) and haplotype networks were made for each cluster using PopArt to aide in morphological validation (Leigh and Bryant 2015). Haplotype networks were coloured according to the plant hardiness zones of the Swedish horticultural society (Fig. 9b) (Riksförbundet Svensk Trädgård 2018).

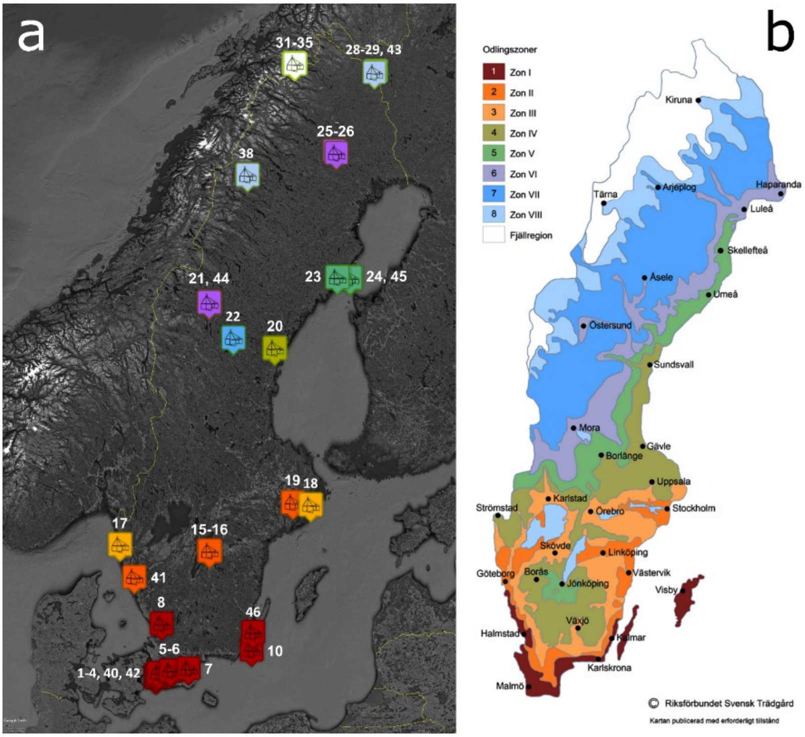


Figure 9. (a) Sites of the Swedish Insect Inventory Project, colour-coded by climatic zones identified by the Swedish Horticultural Society, (b) Climatic zones (odlingszoner) of the Swedish Horticultural Society (Riksförbundet Svensk Trädgård 2018), used with permission).

One-hundred clusters were randomly selected to conduct a thorough morphological validation following previously established characters and character states (Hartop and Brown 2014). In this stage, all main haplotypes (containing at least 20% of cluster specimens) were checked, and then haplotypes across the cluster were checked until no haplotype >1 bp away from a checked cluster remained. Additionally, at least one specimen from each geographic area was checked in this stage. Clusters where morphological results agreed with the initial molecular delimitation were considered validated, while multi- or partial-species clusters failed validation. The set of 100 clusters were then evaluated to determine if properties of clusters were predictive of cluster failure. We tested six explanatory variables: "haplo" (number of haplotypes in a cluster), "spec" (number of specimens in a cluster), "stability" (a measure of cluster stability over thresholds of 1-3%), "max_p" (maximum pairwise distance within a cluster), "zones" (number of geographic zones represented in a cluster), and "sites" (number of sites represented within a cluster). We fitted a generalised linear model with quasibinomial errors to the validation results to assess which of the variables best predicted cluster congruence with morphology.

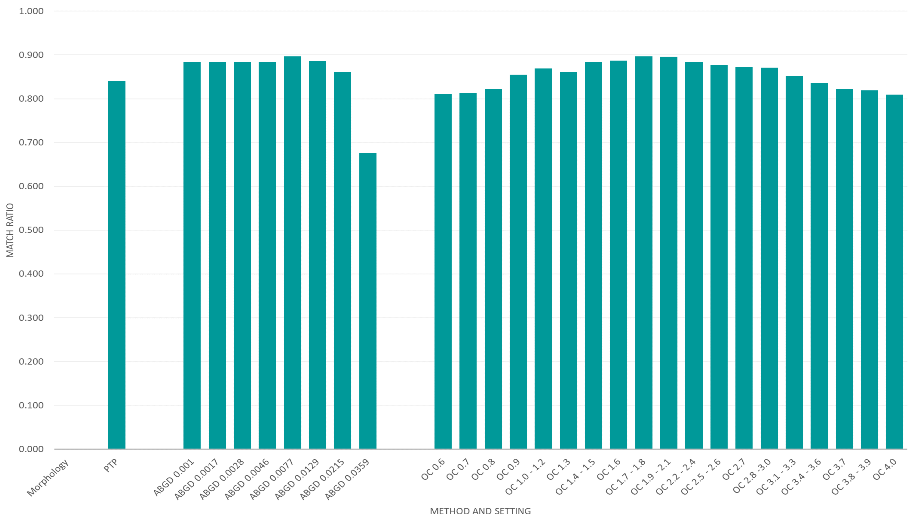


Figure 10. Match ratios for PTP, ABGD (all priors) and OC (all thresholds) versus morphology across methods and settings.

Based on the results from the first 100 test clusters, we used maximum p-distance and stability to flag remaining clusters as potentially incongruent (PI) if they had over 1.5% intracluster distance or if they had instability between 1-3%. The PI designated clusters were evaluated against a test set of non-PI clusters to determine if variables "max_p" and "stability" accurately identified clusters that are incongruent. These two sets were evaluated using a reduced checking scheme that checked main and most distant haplotypes, but eliminated the need to check more specimens across cluster variation or geographic zones, as we had not found these procedures to be informative of cluster failure. The remaining clusters (all non-PI) were evaluated using most distant haplotypes only.

To assess the robustness of our results to variations in the clustering algorithm, we evaluated the match ratio (Ahrens et al. 2016) of our morphologically validated results with results from a range of clustering methods and thresholds: Objective Clustering (OC) from 0.6-4% (Meier et al. 2006), Automated Barcode Gap Discovery (ABGD) across the range of priors (0.001-0.0359) (Puillandre et al. 2012a), and Poisson Tree Process (PTP) (Zhang et al. 2013). We additionally compared the subset of our data that matched Barcode Index Numbers (BINs) in the Barcode of Life Data Systems (BOLD) to indirectly assess results obtained with the Refined Single Linkage (RESL) algorithm that is not published and could therefore not be evaluated directly.

We finally tested the robustness of our methods by pulling all available phorid barcodes from public databases (GenBank and BOLD) and adding them to our dataset. We then evaluated how many barcodes fit into our original Swedish clusters, and how the addition of these clusters shifted clusters from non-PI to PI or resulted in fusion events.

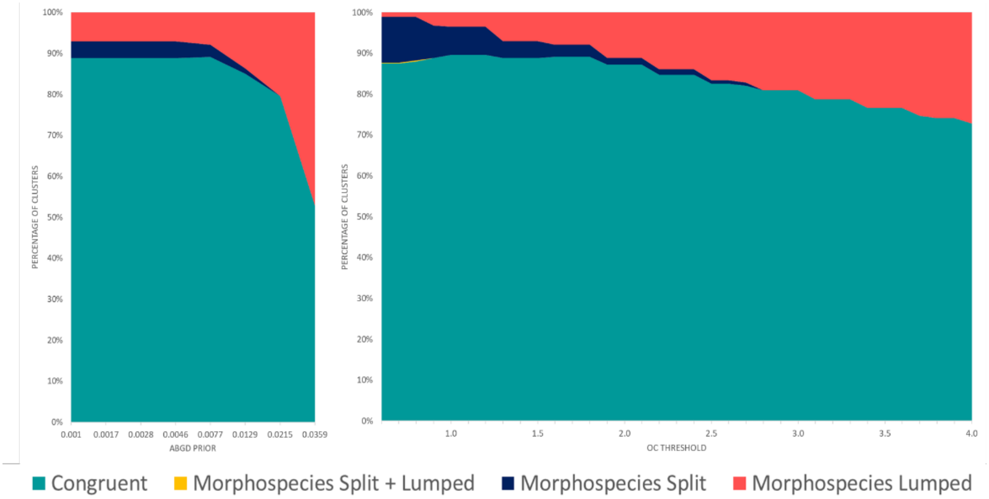


Figure 11. The splitting and lumping of morphological clusters with ABGD (left) and OC (right) across settings.

Results

A total of 19 570 phorid flies were sequenced and the final analysis focused on 17 443 barcodes that grouped into 315 putative species at 3% OC. Evaluation of the first 100 test clusters revealed that 7/100 contained multiple species and an additional two were of uncertain composition ("species complexes"). A generalised linear model revealed that "stability" was the only factor that was significant in predicting cluster failure after removing strongly covarying factors from the model. However, this variable was highly collinear with variable "max_p", so that it, too, would be predictive of cluster failure. We used these two variables to identify clusters as PI in the next stage.

The next stage evaluated 43 PI clusters against 43 non-PI clusters to assess whether the identified cluster properties correctly identified cluster failure. Of the 43 PI clusters, 26% (11/43) were found to contain multiple species, while none of the non-PI clusters did. The evaluation of the remaining 129 smaller, non-PI clusters revealed no further incongruence.

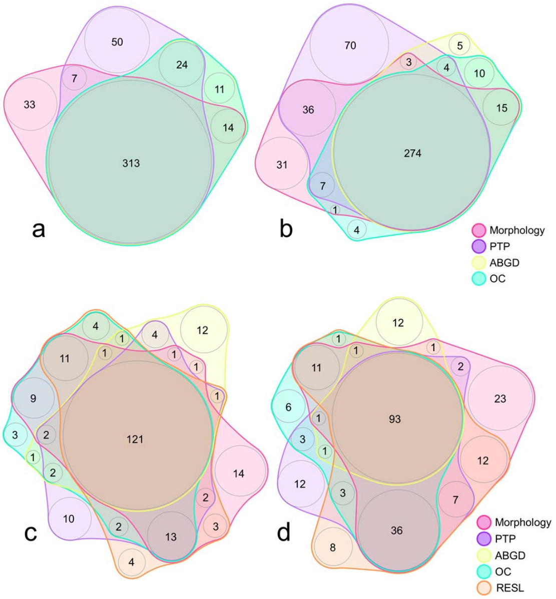


Figure 12. An illustration of the congruence between morphology, PTP, ABGD, and OC methods with (a) optimal settings (ABGD P=0.0077, OC 1.7%) and (b) conservative settings (ABGD P=0.0215, OC 3.0%) and between morphology, PTP, ABGD, OC and RESL methods with (c) optimal settings (ABGD P=0.0077, OC 1.7%) and (d) conservative settings (ABGD P=0.0215, OC 3.0%).

Evaluation of different clustering methods and thresholds revealed that match ratios for OC and ABGD with morphology were maximized at 0.897 (OC 1.6-17%, ABGD 0.0077) (Fig. 10). PTP had a maximum match ratio of 0.841 (Fig. 10). We further evaluated congruence by looking at where methods split and/or lumped species. OC starts lumping morphospecies at 0.6% and stops splitting morphospecies at 2.8%, while ABGD lumps morphospecies across all priors, and stops splitting at p=0.0215 (Fig. 11). PTP both splits and lumps morphospecies. Evaluation of congruence between methods revealed that ABGD and OC were largely congruent, while PTP was an outlier and morphology often revealed species not seen in any of the molecular delimitations (Fig. 12a, b). Similar results were obtained with the subset of specimens evaluated with RESL (Fig. 12c, d).

The addition of barcodes from public databases largely did not affect the original Swedish clusters, as out of the entire dataset of 84 656 barcodes, 58 572 did not match Swedish mOTUs. The clusters that did match, however, affected 244 of the original 329 clusters. There were numerous shifts of clusters to a PI designation and there were also fusion events involving 27 of the original clusters. The overall percentage of PI clusters went from 21% in the original dataset to 43% in the expanded dataset.

Final LIT protocol is a set sequence of steps that (1) identifies potentially problematic clusters (2) selects specimens for validation procedures and (3) follows through with species delimitation for multi-species clusters (Fig. 13).

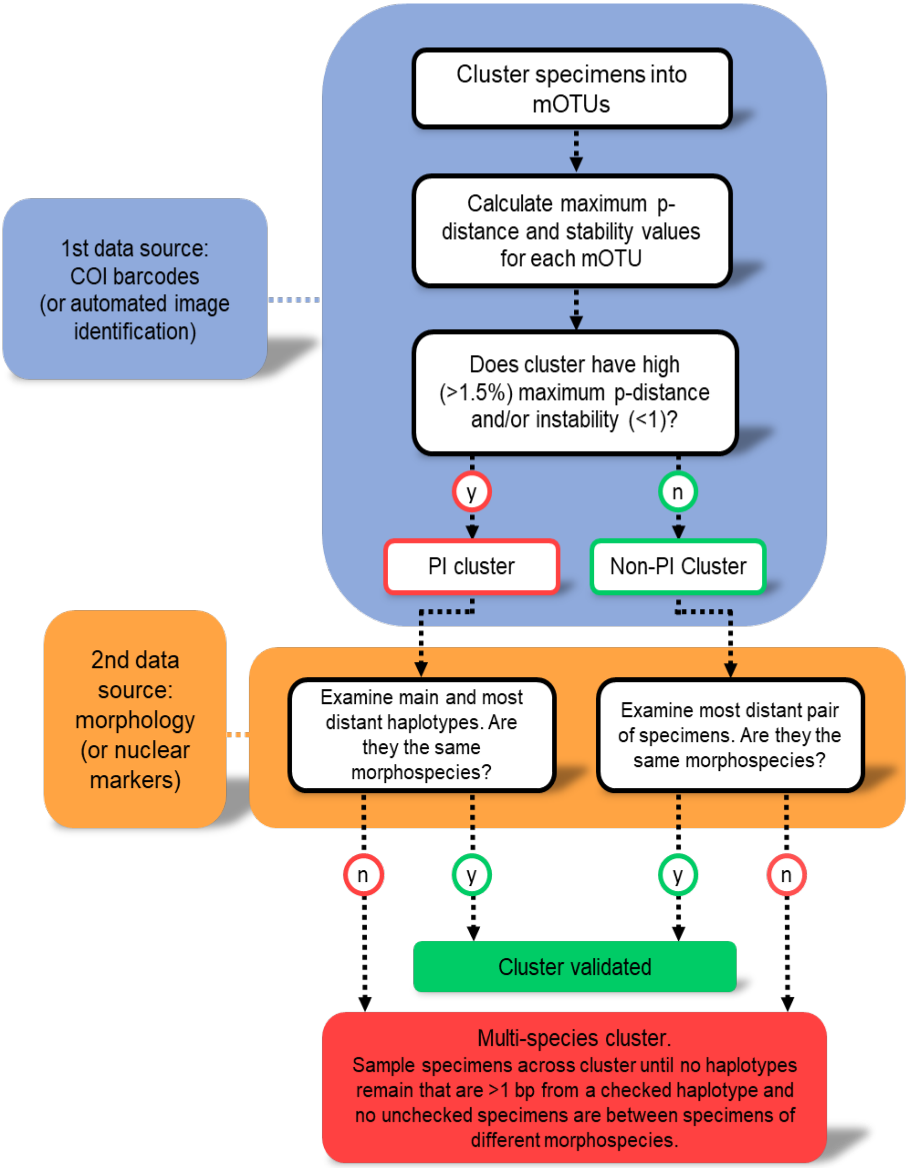

1st data source: COI barcodes (or automated image identification)
Cluster specimens into mOTUs
Calculate maximum p-distance and stability values for each mOTU
Does cluster have high (>1.5%) maximum p-distance and/or instability (<1)?
y
n
PI cluster
Non-PI Cluster
2nd data source: morphology (or nuclear markers)
Examine main and most distant haplotypes. Are they the same morphospecies?
Examine most distant pair of specimens. Are they the same morphospecies?
n
y
y
n
Cluster validated
Multi-species cluster.
Sample specimens across cluster until no haplotypes remain that are >1 bp from a checked haplotype and no unchecked specimens are between specimens of different morphospecies.

Figure 13. A flowchart of final LIT protocol.

Paper IV

Materials and Methods

This project used the same sampling scheme as *paper III* (Fig. 9a) but included barcodes from an additional ca. 14 000 specimens sampled from three other time-periods. The total dataset therefore included ca. 32 000 specimens collected across late spring, midsummer, late summer, and offseason samples. For most analyses, barcodes were clustered using objective clustering optimised to approximate species at 1.7%, as in *paper III* results.

Species richness was analysed using Chao1 and raw species accumulation curves plotted with EstimateS (Colwell 2013) and the R package (R Development Core Team) iNEXT (Chao et al. 2014). iNEXT was also used to plot rarefaction curves. Sites were again categorised according to the plant hardiness zones of the Swedish Horticultural Society (Fig. 9b), and these categorizations were used in non-metric dimensional scaling plots (NMDS) to visualise phorid communities across space and time. Comparative analyses were done using different clustering thresholds and using categorisations based on mean annual temperature and mean annual precipitation maps calculated from the normal period 1991-2020, available from the Swedish Meteorological and Hydrological Institute (smhi.se). Analysis of similarities (ANOSIM) and Similarity percentage analysis (SIMPER) tests were run with PRIMER v7 (Clarke and Gorley 2006) to provide quantitative data on the differences between regions and time periods and to assess the significance of these differences.

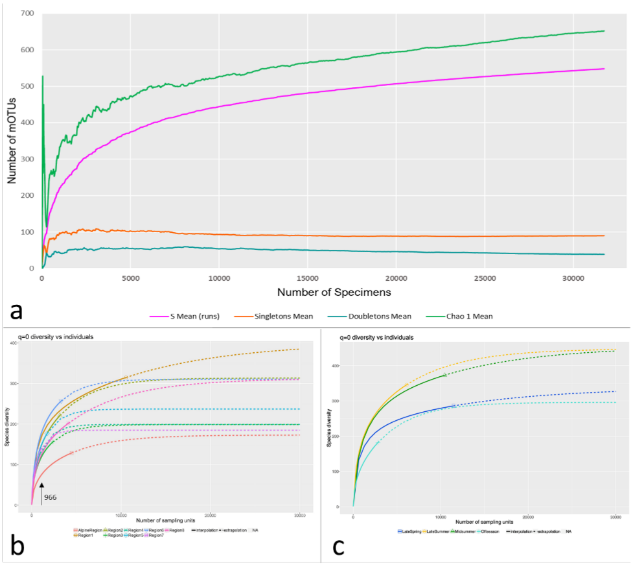


Figure 14. Species accumulation curves including associated Chao1 estimates for (a) Sweden, (b) plant hardiness regions, and (c) time periods, all showing that the fauna is undersampled.

To analyse our data without the imposition of a priori map categorisations, we conducted hierarchal modelling of species communities (HMSC) (Ovaskainen et al. 2017; Ovaskainen and Abrego). We ran four models in total: two with presence-absence data and 2 with abundance data, and one each in both categories using either haplotype or 1.7% mOTUs. We included in these models as fixed effects a single categorical variable representing the four time-periods (late spring, midsummer, late summer, and offseason) ("Time.Period") and five continuous variables: (1) Worldclim mean temperature in the warmest quarter ("bio10") (Fick and Hijmans 2017), (2) Percent forest/woodland cover (50m buffer) ("ForestWood") and (3) Percent agricultural land cover (50m buffer) ("Agriculture"), both from the Swedish National Land Cover Database (http://www.swedishepa.se/State-of-the-environment/Maps-and-map- services/National-Land-Cover-Database/), and (4) number of trapping days per sample ("TrapDays"), and (5) a binary variable indicating whether all specimens in a sample were sequenced or not ("FullSample"). We also included a spatially explicit random effect based on sample site coordinates ("Random: site") and a temporally explicit random effect based on median sampling date ("Random: time").

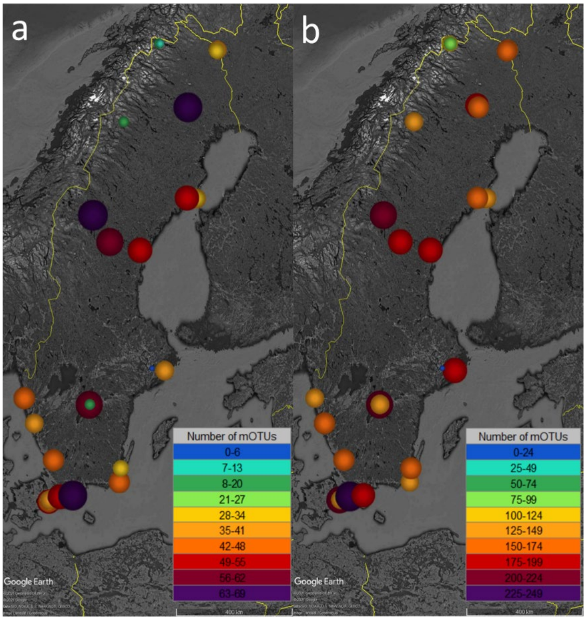


Figure 15. Diversity across sites, (a) Current species richness estimated from samples that were rarefied to make them comparable across sites and (b) Chao1 estimates for each site.

Results

The final dataset consisted of 31 739 specimens that clustered into 549 mOTUs at OC 1.7%. Chao1 species accumulation curves revealed that all sites, regions and time-periods are still undersampled, with a further 100 species awaiting discovery (for a total of 652 estimated species) based on current sampling (Fig. 14). We rarefied current species richness to compare diversity across sites and found no clear pattern in overall richness, with high and low diversity sites found across the country (Fig. 15).

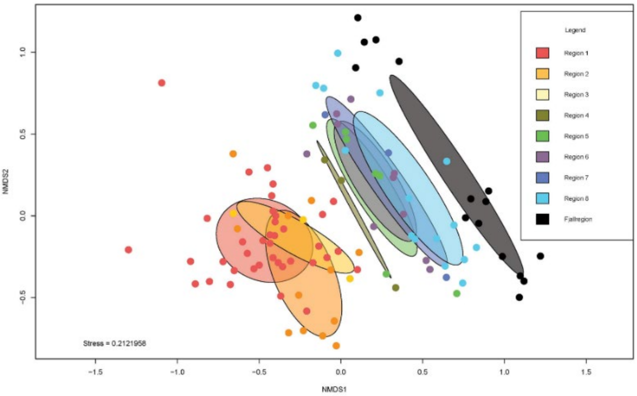


Figure 16. Plot of all samples (1.7% mOTUs, threshold of 100 specimens) colour-coded according to plant hardiness region.

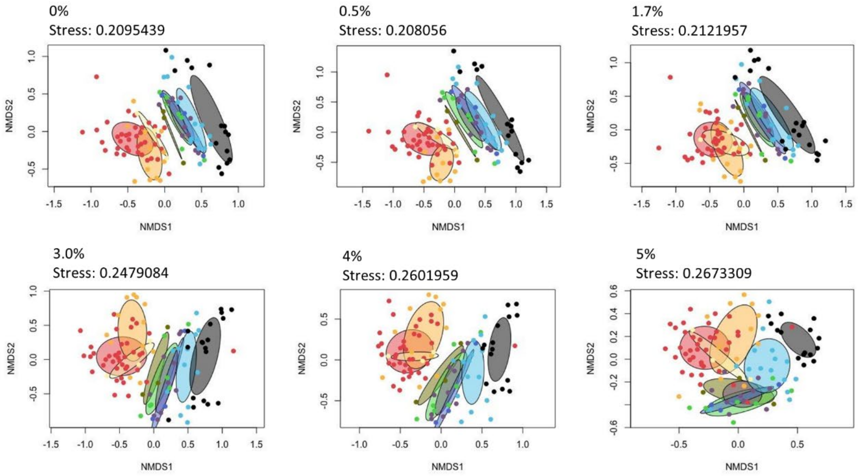


Figure 17. Samples plotted with mOTUs calculated at different thresholds with ellipses coloured according to plant hardiness zones.

Spatial patterns were clearly correlated with the regions of the plant hardiness map visually (Fig. 16) and were arranged in a near linear gradient across Sweden. We confirmed the relationship with a linear regression model of our NMDS1 values plotted against ordered zones, this showed a significant positive linear fit to our data based on either presence-absence data (adjusted R^2=0.79) or abundance data (adjusted R^2=0.67). We additionally analysed this relationship over clustering thresholds from 0% (haplotypes) to 5% and found that patterns were nearly identical from 0-1.7%, after which there was a clear phase shift and then a blurring of patterns as species were lumped together (Fig. 17). We also conducted a comparative analysis using the SMHI temperature and precipitation zones. Visually, the temperature zones were a good fit to our data and closely mimicked the plants hardiness zones (that are largely temperature dependent) and the precipitation map was a poor fit. This was confirmed with linear regression models where the average annual temperature data had an adjusted R^2 of 0.76 and 0.63 with presence-absence and abundance data, respectively, and the precipitation data had an adjusted R^2 of 0.37 and 0.30 with presence-absence and abundance data, respectively.

Seasonal samples were visually apparent linearly across the regional gradient, except for the offseason where the sampling overlapped with other seasons, obscuring patterning (Fig. 18).

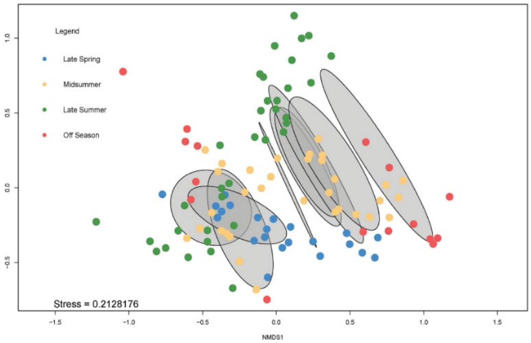


Figure 18. Plot of all samples (threshold of 100 specimens) with regional ellipses greyed out and samples colour coded according to time-period. With the exception of the offseason, seasonal catches run linearly through regional ellipses.

HMSC results confirmed the results that we obtained using categorisations based on a priori maps and additionally showed that models based on either haplotypes or "species equivalent" mOTUs made little difference to results. Proportions of explained variance were largely dependent on temporal (seasonal) factors with >35% of variance explained by variables "Random:time" and "Time.Period" (Fig. 19). Presence-absence models were additionally highly dependent on climate (Fig. 19).

We analysed the positive and negative responses of individual haplotypes and mOTUs to our variables and found that most taxa showed a negative response to offseason trapping in both presence-absence and abundance models, indicating they are less likely to be present and less likely to occur in high numbers, in the winter. Taxon occurrences showed a mixture of positive and negative responses to the temperature variable, reflecting the community gradient across Sweden we observed with our plant region NMDS plots.

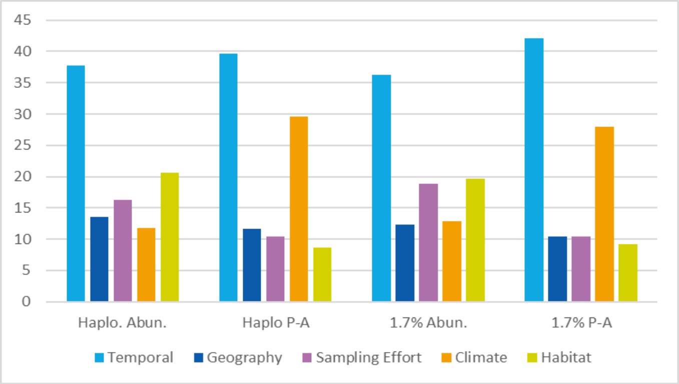


Figure 19. Proportion of explained variance for the four models. All four models are largely temporally driven which includes variables “Random:time” and “Time.Period”. Presence-absence models are also largely driven by the climatic variable “bio10”. Geography is variable “Random: site”, Sampling effort is “FullSample” and “TrapDays”, and Habitat is variables “Agriculture” and “ForestWood”.

Residual correlations between taxa across sites were common for 1.7% mOTUs but less so across time. Haplotype data had many fewer correlations, this may reflect the relative rarity of haplotypes across the dataset. Residual correlations were extremely low in the abundance models, indicating abundance patterns were well modelled by the covariates of our models.

Discussion

Here, we have taken a multi-faceted approach to a dark taxon by gathering data and developing tools to facilitate a better understanding of a diverse and abundant group. This is critically needed in this time of great anthropogenic change when we need to prioritise biodiversity assessments and ground them in quantitative reasoning. Dark taxa represent some of the most species rich and abundant groups of organisms. They represent a wide variety of life histories and a significant proportion of terrestrial animal biomass. We must prioritise understanding these groups. To that end, in this thesis we have first created a phylogenetic framework to guide future studies on *Megaselia* (*paper I*), we have then optimised methods for species discovery and delimitation from large-scale molecular datasets (*papers II* and *III*), and finally, we have utilised these datasets to conduct a countrywide analysis of the richness and spatiotemporal distribution of an enigmatic group.

In *Paper I* we discovered a potentially monophyletic core of *Megaselia.* This gives us hope that a natural circumscription and diagnosis of the genus based on putative morphological apomorphies might be on the horizon. The clades that fall outside the core of *Megaselia* will require further investigation, but our preliminary morphological diagnoses may help to guide delimitations of the genera closely related to *Megaselia*. Within the core of *Megaselia*, we will need to collect more data to give reliable morphological diagnoses to the species groups we have identified. We have confirmed that previous divisions within *Megaselia* were often unnatural and we have provided the framework for alternative divisions of the genus based on phylogenetic data. Studies on the tentative species groups we have identified must proceed utilising both morphological and molecular data to gain a better understanding of these groups and the relationships between them.

The molecular framework for *Megaselia* from *paper I* is far from comprehensive. The 145 species it includes are roughly one-quarter of the species-level molecular units found from Sweden for the dataset used in *paper IV*. Therefore, we must be cautious as it covers only a small fraction of the diversity from Sweden, a Palaearctic country that contains a small part of the world fauna. We will need many more species, additional outgroup representatives, and sampling from many more regions before we get anywhere close to a comprehensive grasp of the evolutionary history of *Megaselia* and the clades within.

Papers II and *III* build on a previously established "reverse workflow" where large numbers of specimens are NGS barcoded and sorted to putative species in preparation for work by specialists (Srivathsan et al. 2018, 2019, 2021; Wang et al. 2018b; Yeo et al. 2020). This process allows for the integration of multiple data sources into taxonomy in an accurate, efficient, and affordable way. Although integrative taxonomy is accepted best practice (Dayrat 2005; Padial and Miralles 2010; Schlick-Steiner et al. 2010; Pante et al. 2015; Vitecek et al. 2017), previous efforts to develop a system for large-scale integrative taxonomy have been underdeveloped (but see Puillandre et al. 2012b; Kekkonen and Hebert 2014). This is largely because barcoding was too expensive to sequence large numbers of specimens. Instead, specimens were often sorted to morphospecies and then representatives were picked for sequencing (Butcher et al. 2012; Riedel et al. 2013b; Lücking et al. 2016). Not only is such a process time consuming, but it is prone to error, especially if morphospecies sorting is conducted by parataxonomists and not by specialists (Krell 2004). Even when experts do the preliminary sorting, this approach cannot detect cryptic species or easily process large numbers of specimens from groups that are exceedingly difficult to determine. Fortunately, we have moved beyond the formal DNA extraction and Sanger sequencing techniques that were too costly to allow the processing of all (or many) specimens in samples. The ability to process large amounts of material offers many advantages over previous approaches. First, specimens are immediately sorted to putative species, allowing the attention of experts to be focused on taxa of interest. An expert working with the reverse workflow receives all specimens pre-sorted, with both sexes and all life stages (if collected) already grouped together. The lab work does not require any specialised skills but instead relies on technicians who can perform extraction and PCR at the rate of ca. 10 microplates (96 wells) per day. The time and expertise required for an NGS reverse workflow barcoding approach to species sorting and discovery is a small fraction of that required for the variety of morphological approaches commonly used today.

In *paper II*, we optimised a method of individual NGS barcoding on the Oxford Nanopore Technologies (ONT) MinION using 1D sequencing. We demonstrated that with an improved bioinformatics pipeline, new primer tags, and re-pooling of weak products it was possible to increase the capacity of a standard MinION flowcell to ~3 500 specimens while reducing costs to <0.35 USD per barcode. The ability to identify weak products in the first run and re-pool for a second run within a day or so is crucial when dealing with thousands of specimens. This can be contrasted with sequencing at centralised facilities, where re-pooling of select DNA extracts for new analysis might not be possible for weeks after the initial run. The workflow presented in *paper II* made it feasible to use inexpensive mobile technology for largescale barcoding. Since the publication of that paper in 2019, the MinION has undergone a number of advancements and, with continued work with this technology, we are now able to sequence up to 10 000 specimens per flowcell (Srivathsan et al. 2021). The optimisations of such workflows can provide inexpensive and efficient solutions to species discovery and delimitation in dark taxa. This is greatly expanded upon in *paper III* where minibarcodes (313 bp) are used as the first data source in the delimitation of 365 species. Although Yeo et al. (2020) have shown that minibarcodes perform comparably to traditional 658 bp barcodes for the delimitation and identification of species, smaller fragment lengths were necessitated by sequencing platforms like Illumina. Now that MinION is a reasonable alternative, full length barcodes can easily be swapped out in workflows like the Large-scale Integrative Taxonomy (LIT) proposed in *paper III.*

Working with portable sequencers is important as the world becomes increasingly focused on democratising science. The MinION is a good low-cost option for those without access to centralised molecular facilities. It requires only basic molecular lab equipment and computing power and can therefore become a tool to bring molecular sequencing to countries that lack the resources for molecular sequencing, but also into grade schools and the homes of citizen/community scientists.

The last goal of *paper II* was the first use of a system that guides well-informed species descriptions. As taxonomists have struggled to streamline species descriptions, various forms of “turbotaxonomy” have been suggested. Often, these systems involve the use of molecular barcodes in combination with photographs and reduced morphological descriptions (Butcher et al. 2012; Riedel et al. 2013b). We decided to take the approach of barcodes in combination with photographs, but without written morphological descriptions. This requires that the photographs be of sufficient quality for a reader to be able to observe in them any characters that normally would have been expressed in text. We do include a written diagnosis to point out salient features. A similar approach, using barcodes and morphology without morphological descriptions, has recently been applied to wasps in the family Braconidae (Meierotto et al. 2019; Sharkey et al. 2021). A key difference between the two approaches is that of validation. In our description of *Megaselia sepsioides* sp. nov., we not only document morphological variation with high quality photographs, but we also document molecular variation with a haplotype network of all specimens we had available for description. We consider these haplotype networks to be a crucial, formal addition to our version of a turbotaxonomic description. Documenting all variation observed in a species must be done as clearly as possible to help inform future work. As data on more species and specimens are added to our body of knowledge, we will be able to separate species from intraspecific variation utilising all available data. The processes of discovery, delimitation and validation for *M. sepsioides* inspired a great expansion of these methods in *paper III*.

In *paper III* we made the processes of species discovery and delimitation explored in *paper II* systematic and objective by formalising procedures for moving from samples to large-scale barcode datasets to validated species units. We developed the Large-scale Integrative Taxonomy (LIT) system to transform the study of dark taxa. We showed that we can accelerate biodiversity discovery while integrating multiple data sources by obtaining a first data source for large numbers of specimens and then systematically and objectively targeting a subset of specimens for the acquisition of a second, more expensive data source. In our study, we first NGS barcoded ca. 18 000 phorid specimens and then obtained morphological data for a small fraction of these (final protocol would require the checking of 915/18 000 specimens), but in the future a first data source could be automated image identification and a second (or even third) data source could be nuclear markers or ecological data.

To reduce the number of specimens that required examination while guaranteeing that the validation process was thorough, we first identified how to flag clusters that were most likely to be incongruent with morphology. This allowed us to develop a final protocol where the minimum number of specimens (2 from the most distant haplotypes) can be checked from clusters that are not flagged as "potentially incongruent", while clusters that are more likely to be problematic can be checked with 5-7 specimens from main and most distant haplotypes. The two variables that flag clusters for deeper examination are logical, but efficient. Maximum p-distance over 1.5% identifies clusters that have high amounts of variation, while stability identifies clusters that have splits that would occur at thresholds below those used for clustering. In future, a potential alternative to stability values (that require a hierarchical structuring of clusters) might be the maximum branch length of a median joining network for a cluster.

The LIT protocol may be subject to complications with expanded sampling and geographic scope. Although many of the public database barcodes did not fit into our original clusters, those that did shifted many clusters to a PI designation and caused several fusion events. Although these data were downloaded from public databases and we therefore have a limited ability to control for quality and no ability to check morphology, they offer a look ahead at the potential issues the may be faced when moving beyond regional studies. It will be impossible to know exactly how to accommodate expansion problems until they occur. Clustering at lower thresholds as interspecific distances decrease is one option, but we will not know how well this will work until we have much larger datasets to work on, and access to specimens for validation. Changing the second data source (from morphology to nuclear markers) or expanding to a sequential workflow that uses a third data source for challenging areas are both options. LIT may also turn out to be best applied to regional datasets, and not to a worldwide fauna.

Our study in *paper III* clearly shows that no method or threshold of molecular clustering is accurate enough to be used without validation with a second data source. Even at optimal settings, there are a number of morphospecies that are not detected by any clustering method. This is not surprising, as algorithms are most likely to delimit improperly (although sometimes in agreement with each other) in areas where there are large, rapid radiations (Puillandre et al. 2012a; Ratnasingham and Hebert 2013; Zhang et al. 2013). In such cases, integrative taxonomy will be a necessity to properly identify species. Our study revealed that such complex areas may be a large proportion of some taxa. Of the 17 443 specimens in our final dataset, 41% (7 150) belonged to 3% OC clusters that contained multiple species. If these specimens were identified based on barcode data only, most would be misidentified.

LIT is a precursor to species description, as conducted in *paper II* for *M. sepsioides*. In future, the specimens from *paper III* will go through a similar workflow. The work to describe *Megaselia sepsioides* was done in a day but in the future, when the many species from *papers III* (and *IV*) need to go through the same pipeline, the descriptive process (photography and the creation of haplotype networks) must be further optimised. This can be done by automating photography and the creation of haplotype networks for several species per day. Even if just five species could be handled per day, the description of all species analysed in *paper III* would take under a year. On the other hand, sorting, discovering, and describing as many specimens and species with traditional methods would take years, perhaps decades. This is something of a moot point, however, because such work is rarely completed for hyperdiverse groups like *Megaselia*. More often, groups like *Megaselia*, and phorids more generally, are left unsorted in bulk samples after other (usually larger, more charismatic) taxa have been sorted out. Thus, the name "dark taxa".

The investigation of some fundamental questions about Phoridae, as a representative dark taxon, is the subject of *Paper IV*. Unlike better studied groups, for dark taxa we often do not have knowledge of fundamental questions regarding species richness and distributions. Our investigations into the phorids of Sweden revealed an estimated 652-713 species (based on Chao 1 and CNE estimates as upper and lower limits) (CNE from Ronquist et al. 2020). This is well beyond the 374 species currently known from the country and reveals that there is still much to explore, even in a country where the biodiversity is as well-known as in Sweden. It is not surprising to have estimates of scuttle fly diversity that may be close to twice the known fauna – a study in Los Angeles, CA documented up to 82 species living in backyards, and a suburban garden in Cambridge, UK has been documented to contain nearly 100 species (Brown and Hartop 2016).

Future efforts to capture the remaining fauna can focus on filling in geographic or habitat gaps using baseline material from the Swedish Malaise Trap Project (Karlsson et al. 2020) or newer material from the ca. 200 Malaise trap sites of the Insect Biome Atlas project (http://insectbiomeatlas.org). Additionally, the material thus far examined has been exclusively from Malaise traps. Utilising other trapping methods would undoubtedly reveal new species and raise faunal estimates. Even still, the diversity across Sweden seems unlikely to exceed the species diversity captured by the single trap in *paper II* that gave an estimate of >1 000 species at that site based on far less sampling. The extrapolation of this information to an estimate of 100 000 species of *Megaselia* in the Afrotropical region will certainly be revised as we gather more data and better understand species turnover. *Paper IV* reveals that phorid communities exist in spatial and temporal gradients, and that nearly all the dissimilarity of phorid communities of Sweden is due to turnover rather than nesting of species. This makes the Afrotropical estimate seem likely to be a reasonable one. Previous speculation had often centred around guesses of 10 000-20 000 species of *Megaselia* worldwide, but data from *papers II* and *III* suggest these may be significant underestimates. This is especially true considering that BLAST results from *paper II* did not have even one species-level match to the ca. 80 000 *COI* sequences in NCBI's GenBank. Further sampling and analysis will be needed to determine if there is truly no overlap between the phorids found at a site in Uganda and the phorids currently sequenced from the rest of the world.

Another focus of *paper IV* was whether spatial and temporal studies could be conducted using haplotype data. This may seem a strange focus after all the work in *paper III* to determine species units, but it is important to understand where dark taxa can be studied without taxonomic interpretation, and where specialist attention will be critical to results. The revelation that haplotype data yielded comparable results to “best species” approximation mOTUs in our community ordinations is promising news for molecular ecologists interested in patterns across time and space, and in response to change.

Our hierarchical species modelling (Ovaskainen et al. 2017; Ovaskainen and Abrego 2020) also revealed similar results with haplotypes and species proxy mOTUs. HMSC showed that phorid communities are explained by temporal (seasonal) variables, and that presence/absence is additionally driven by climate (in a gradient across latitude and altitude). Some species showed positive responses to woodland/forest or agricultural habitats, while others did not. However, our models clearly captured only some of the important factors, as residual plots of species associations reveal that a large part of the variation is due to missing unknown variables. We hypothesise that microclimate associated with soil, or variable within the soil, may turn out to be important to explain the structuring of phorid communities, and perhaps also many other dark taxa, as many exist just above, or sometimes within, the soil. We will explore further variables in future modelling, and we also hope to incorporate phylogenetic information to determine whether observed patterns have an evolutionary basis.

In additional to their tremendous species and ecological diversity and abundance, our results from *paper IV* suggest that dark taxa may offer a fine-scale approach to biodiversity studies. Although sampling is limited thus far, there is potential for regional or even local endemism based on currently observed distributions. Even at a larger scale, dark taxa may offer patterns of distribution that are different from more commonly studied taxa like birds and butterflies. We downloaded bird and butterfly data from standard sampling schemes in Sweden and found that for both groups, approximately 2/3 of species are widely distributed across the country. In contrast, based on current sampling, nearly 2/3 of phorids species are restricted to the northern or southern part of the country. Admittedly, our phorid results are based on an undersampled fauna and the sampling schemes for birds and butterflies are quite different, utilising many more sites/transects over shorter periods of time. Until we have the phorids of Sweden much more completely sampled and a better comparison can be made, we will not know for sure how unique their patterns of distribution may be.

Conclusion

Dark taxa require a unique approach, but they may yield results that are impossible to obtain with taxa that are easier to study. Due to overwhelming species and specimen numbers combined with challenging taxonomy, a multi-faceted approach is essential. Here, we combined several approaches to better understand the fly family Phoridae by first constructing a framework on which to base future studies (*paper I*), then improving methods for large-scale species discovery and delimitation (*papers II* and *III*), and finally expanding knowledge of how phorid communities change across time and space (*paper IV*). These are all significant advances in our knowledge of Phoridae, and especially of the genus *Megaselia*, which makes up a majority of specimens and species in temperate samples. Widespread adoption and further refinement of such methods to study other dark taxa will rapidly advance our understanding of the world's biodiversity. This will be facilitated by developments in image recognition, robotics, and sequencing technologies, all of which will make the study of dark taxa increasingly inexpensive and efficient.

With ever increasing societal concern over the implications of biodiversity loss for the long-term survival of mankind, this is urgent work and there is still much to be done. We must be able to accurately document our biodiversity as it undergoes rapid losses and transformations due to climate change and habitat loss. To do this, we must focus on taxa that contribute great species richness, significant biomass, and great biological diversity. Doing so would result in a truly quantitative approach to biodiversity studies that is much needed to accurately assess our natural systems and is especially critical in this time of great anthropogenic change.

References

Aberer A.J., Krompass D., Stamatakis A. 2013. Pruning Rogue Taxa Improves Phylogenetic Accuracy: An Efficient Algorithm and Webservice. Syst. Biol. 62:162–166.

Ahrens D., Fujisawa T., Krammer H.-J., Eberle J., Fabrizi S., Vogler A.P. 2016. Rarity and Incomplete Sampling in DNA-Based Species Delimitation. Syst. Biol. 65:17.

Brown B.V., Borkent A., Adler P.H., De Souza Amorim 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., Hash J., Hauser M., Hippa H., Ibáñez-Bernal S., Jaschhof M., Kameneva E.P., Kerr P.H., Korneyev V., Korytkowski C.A., Kung G., Kvifte G.M., Lonsdale O., Marshall S.A., Mathis W., Michelsen V., Naglis S., Norrbom A.L., 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., Whitworth 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. 1:8.

Brown B.V., Hartop E.A. 2016. Big data from tiny flies: patterns revealed from over 42,000 phorid flies (Insecta: Diptera: Phoridae) collected over one year in Los Angeles, California, USA. Urban Ecosyst.

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

Chao A., Gotelli N.J., Hsieh T.C., Sander E.L., Ma K.H., Colwell R.K., Ellison A.M. 2014. Rarefaction and extrapolation with Hill numbers: a framework for sampling and estimation in species diversity studies. Ecol. Monogr. 84:45–67.

Clarke K., Gorley R.N. 2006. PRIMER v6: user manual/tutorial. PRIMER-E, Plymouth. 29:1060–1065.

Colwell R.K. 2013. EstimateS: Statistical estimation of species richness and shared species from samples. Version 9. User's Guide and application published at: http://purl.oclc.org/estimates.
.

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

Disney R.H.L. 1979. Natural history notes on some British Phoridae (Diptera) with comments on a changing picture. Entomol. Gaz. 30:141–150.

Disney R.H.L. 1989. A key to Australasian and Oriental Woodiphora (Diptera: Phoridae), affinities of the genus and description of new species. J. Nat. Hist. 23:1137–1175.

Disney R.H.L. 1990. Some myths and the reality of scuttle fly biology. Antenna. 14:64–67.

Disney R.H.L. 1994. Scuttle flies: the Phoridae. London: Chapman and Hall.

Enderlein G. 1924. Zur Klassifikation der Phoriden und ber vernichtende Kritik. Entomol. Mitteilungen. 13:270–281.

Fick S.E., Hijmans R.J. 2017. WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas. Int. J. Climatol. 37:4302–4315.

Goloboff P., Catalano S. 2016. TNT Version 1.5, including a full implementation of phylogenetic morphometrics. Cladistics. 32.

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

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

Karlsson D., Hartop E.A., Forshage M., Jaschhof M., Ronquist F. 2020. The Swedish Malaise Trap Project: A 15 Year Retrospective on a Countrywide Insect Inventory. Biodivers. Data J. 8:e47255.

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

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

Krell F.-T. 2004. Parataxonomy vs. taxonomy in biodiversity studies – pitfalls and applicability of 'morphospecies' sorting. Biodivers. Conserv. 13:795–812.

Kurina O. 2012. Description of four new species of Zygomyia Winnertz from Ethiopia and Uganda (Diptera: Mycetophilidae). Afr. Invertebr. 53:205–220.

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

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

Marshall S.A. 2012. Flies: The Natural History and Diversity of Diptera. Buffalo, New York: Firefly Books.

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

Meierotto S., Sharkey M.J., Janzen D.H., Hallwachs W., Hebert P.D.N., Chapman E.G., Smith M.A. 2019. A revolutionary protocol to describe understudied hyperdiverse taxa and overcome the taxonomic impediment. Dtsch. Entomol. Z. 66:119–145.

Ovaskainen O., Abrego N. 2020. Joint Species Distribution Modelling. Cambridge University Press.

Ovaskainen O., Tikhonov G., Norberg A., Guillaume Blanchet F., Duan L., Dunson D., Roslin T., Abrego N. 2017. How to make more out of community data? A conceptual framework and its implementation as models and software. Ecol. Lett. 20:561–576.

Padial J.M., Miralles A. 2010. The integrative future of taxonomy. :14.

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. 2015. DNA barcoding and taxonomy: dark taxa and dark texts. Preprint.:1–13.

Pante E., Schoelinck C., Puillandre N. 2015. From Integrative Taxonomy to Species Description: One Step Beyond. Syst. Biol. 64:152–160.

Puillandre N., Lambert A., Brouillet S., Achaz G. 2012a. ABGD, Automatic Barcode Gap Discovery for primary species delimitation: ABGD, AUTOMATIC BARCODE GAP DISCOVERY. Mol. Ecol. 21:1864–1877.

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

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

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

Riedel A., Sagata K., Surbakti S., Rene T., Michael B. 2013b. One hundred and one new species of Trigonopterus weevils from New Guinea. Zookeys.:1–150.

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

Ronquist F., Forshage M., Häggqvist S., Karlsson D., Hovmöller R., Bergsten J., Holston K., Britton T., Abenius J., Andersson B., Buhl P.N., Coulianos C.-C., Fjellberg A., Gertsson C.-A., Hellqvist S., Jaschhof M., Kjærandsen J., Klopfstein S., Kobro S., Liston A., Meier R., Pollet M., Riedel M., Roháček J., Schuppenhauer M., Stigenberg J., Struwe I., Taeger A., Ulefors S.-O., Varga O., Withers P., Gärdenfors U. 2020. Completing Linnaeus's inventory of the Swedish insect fauna: Only 5,000 species left? PLOS ONE. 15:e0228561.

Ronquist F., Teslenko M., van der Mark P., Ayres D.L., Darling A., Höhna S., Larget B., Liu L., Suchard M.A., Huelsenbeck J.P. 2012. MrBayes 3.2: Efficient Bayesian Phylogenetic Inference and Model Choice Across a Large Model Space. Syst. Biol. 61:539–542.

Schlick-Steiner B.C., Steiner F.M., Seifert B., Stauffer C., Christian E., Crozier R.H. 2010. Integrative Taxonomy: A Multisource Approach to Exploring Biodiversity. Annu. Rev. Entomol. 55:421–438.

Schmitz H. 1953. Phoridae. In: Lindner E., editor. Die Fliegen der palaearktischen Region. Stuttgart: E. Schweizerbart'sche Verlagsbuchhandlung. p. 273–320.

Schmitz H. 1955. Phoridae. In: Lindner E., editor. Die Fliegen der palaearktischen Region. Stuttgart: E. Schweizerbart'sche Verlagsbuchhandlung. p. 321–386.

Schmitz H. 1956. Phoridae. In: Lindner E., editor. Die Fliegen der palaearktischen Region. Stuttgart: E. Schweizerbart'sche Verlagsbuchhandlung. p. 369–416.

Schmitz H. 1957. Phoridae. In: Lindner E., editor. Die Fliegen der palaearktischen Region. Stuttgart: E. Schweizerbart'sche Verlagsbuchhandlung. p. 417–464.

Schmitz H. 1958. Phoridae. In: Lindner E., editor. Die Fliegen der palaearktischen Region. Stuttgart: E. Schweizerbart'sche Verlagsbuchhandlung. p. 465–512.

Schmitz H., Beyer E. 1965a. Phoridae. In: Lindner E., editor. Die Fliegen der palaearktischen Region. Stuttgart: E. Schweizerbart'sche Verlagsbuchhandlung. p. 513–560.

Schmitz H., Beyer E. 1965b. Phoridae. In: Lindner E., editor. Die Fliegen der palaearktischen Region. Stuttgart: E. Schweizerbart'sche Verlagsbuchhandlung. p. 561–608.

Schmitz H., Delage A. 1974. Phoridae. In: Lindner E., editor. Die Fliegen der palearktischen Region. Stuttgart: E. Schweizerbart'sche Verlagsbuchhandlung. p. 638–664.

Schmitz H., Delage A. 1981. Phoridae. In: Lindner E., editor. Die Fliegen der palearktischen Region. Stuttgart: E. Schweizerbart'sche Verlagsbuchhandlung. p. 665–712.

Sharkey M.J., Janzen D.H., Hallwachs W., Chapman E.G., Smith M.A., Dapkey T., Brown A., Ratnasingham S., Naik S., Manjunath R., Perez K., Milton M., Hebert P., Shaw S.R., Kittel R.N., Solis M.A., Metz M.A., Goldstein P.Z., Brown J.W., Quicke D.L.J., van Achterberg C., Brown B.V., Burns J.M. 2021. Minimalist revision and description of 403 new species in 11 subfamilies of Costa Rican braconid parasitoid wasps, including host records for 219 species. ZooKeys. 1013:1–665.

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

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. .

Srivathsan A., Lee L., Katoh K., Hartop E., Narayanan Kutty S., Wong J., Yeo D., Meier R. 2021. MinION barcodes: biodiversity discovery and identification by everyone, for everyone. BioRxiv Prepr.

Stamatakis A. 2014. RAxML Version 8: A tool for Phylogenetic Analysis and Post-Analysis of Large Phylogenies. Bioinformatics.

Swiss Re Institute. 2020. Biodiversity and Ecosystem Services: A business case for re/insurance. .

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

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

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

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

Yeo D., Srivathsan A., Meier R. 2020. Longer is not always better: Optimizing barcode length for large-scale species discovery and identification. Syst. Biol.:syaa014.

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