1D MinION sequencing for large-scale species discovery: 7000 scuttle flies (Diptera: Phoridae) from one site in Kibale National Park (Uganda) revealed to belong to >650 species

Amrita Srivathsan[1], Emily Hartop[2,5,6], Jayanthi Puniamoorthy[1], Wan Ting Lee[1], Sujatha Narayanan Kutty[1,4], Olavi Kurina[3], Rudolf Meier[1,4*]

[1] Department of Biological Sciences, National University of Singapore, 14 Science Drive 4

[2] Zoology Department, Stockholms Universitet, Stockholm, Sweden

[3] Estonian University of Life Sciences, Kreutzwaldi 5D, Tartu, Estonia

[4] Tropical Marine Science Institute, National University of Singapore, Singapore

[5] Station Linné, Öland, Sweden

[6] Naturhistoriska Riksmuseet, Stockholm, Sweden

***Corresponding author**: Rudolf Meier: meier@nus.edu.sg

Other email addresses:

Amrita Srivathsan: asrivathsan@gmail.com

Emily Hartop: emhartop@gmail.com

Jayanthi Puniamoorthy: dbsjay@nus.edu.sg

Wan Ting Lee: dbslwt@nus.edu.sg

Sujatha Narayanan Kutty: tmssunk@nus.edu.sg

Olavi Kurina: Olavi.Kurina@emu.ee

Keywords: NGS barcoding, DNA barcoding, Nanopore sequencing, MinION, large-scale species discovery

ABSTRACT

Background: Most animal species remain to be discovered. We recently proposed to tackle this problem using a ‘reverse workflow’ where all specimens are barcoded via tagged amplicon sequencing and then sorted into putative species (mOTUs). We furthermore suggested that the COI barcodes can be obtained with minimal laboratory equipment using MinION sequencing, but our test with 1D^2 reads only yielded 500 barcodes per flowcell.

Results: Here we show how MinION 1D sequencing can be used to obtain ~3500 COI barcodes per flowcell. Based on 7062 MinION barcodes for a hyper-diverse family of flies (Diptera: Phoridae) collected by one trap in Kibale National Park, Uganda, we discover ~650 species which exceeds the number of phorid species described for the entire Afrotropical region. Our updated MinION pipeline increases processing speed via parallelization, improves demultiplexing, and yet yields reliable barcodes (99.99% accuracy) and similar mOTUs as Illumina sequencing (match ratio: 0.989). Morphological examination of 100 mOTUs confirms good congruence (93% of mOTUs; >99% of specimens). Nearly 90% of species and specimens belong to the megadiverse genus *Megaselia* which is routinely neglected because its species diversity and abundance is too overwhelming. We show that it can be tackled with the reverse workflow. We also illustrate how the molecular data guides the description of a new species: *Megaselia sepsioides* sp. nov..

Conclusions: MinION is suitable for reliable, rapid, and large-scale species discovery in hyperdiverse taxa. Approximately 3500 specimens can be sequenced using one MinION flowcell at a barcode cost of <0.35 USD.

INTRODUCTION

Most life on earth has yet to be discovered with an estimated 80% of extant species being still unknown to science [1]. The majority of these species belong to hyper-diverse and species-rich invertebrate clades. These taxa are ubiquitous, contain most of the multicellular animal species, and often occur in great abundance. However, new species are difficult to find and delimit because it requires the study of thousands of specimens. Typically, this process starts with sampling specimens with bulk trapping methods (e.g. Malaise trap, fogging, pitfall traps, flight intercept traps). It usually yields thousands of specimens per site that need to be sorted: first to higher-level taxonomic groups by parataxonomists and then to species-level by taxonomic experts. The latter work has to be carried out by taxonomic experts because species-level sorting by parataxonomists tends to yield unreliable results [2]. Once morpho-species have been obtained, they are then often tested via DNA barcodes (658 bp fragment of COI) by sequencing a few representative specimens for each morpho-species [3]. This traditional workflow works well for taxa with small numbers of species and specimens but is so time-consuming for hyperdiverse and abundant clades that they are neglected. This is partially responsible for our lack of baseline data for many insect taxa. The traditional workflow has the additional downside that morphologically cryptic species are overlooked. This is particularly likely to happen when expensive Sanger sequencing is used because only few specimens can be barcoded and the probability of detecting cryptic species is low [4].

An alternative approach to species discovery is the ‘reverse workflow’ where every specimen in a sample is individually sequenced (or barcoded) without the destruction of the specimen [4-6]. The specimens are then grouped to molecular Operational Taxonomic Units (mOTUs) based on DNA barcodes. The morphological check of the putative species delimited with DNA barcodes comes last. The taxonomic expert works on pre-sorted material and rectifies mis-sorted specimens, identifies known species, and describes new species. This would have been deemed unrealistically expensive prior to the advent of the High Throughput Sequencing technologies. However, sequencing platforms like Illumina and PacBio are sufficiently cost-effective and are now replacing barcoding via expensive Sanger sequencing [4, 5, 7-10]. For example, sequencing tens of thousands of specimens with Illumina HiSeq can cost as little as 0.17 USD per specimen (including PCR cost, see discussion in Wang et al., 2018 [4]). However, Illumina and PacBio sequencing have some downsides. They are only cost-effective if >10,000 specimens have to be barcoded, sequencing usually has to be outsourced (i.e., the amplicons have to be shipped to a sequencing facility), and it often takes weeks to obtain data. It would be desirable to have alternatives that are fast, scalable, and yet cost-effective. This would be particularly useful if barcoding has to be accomplished under field conditions or in countries with limited access to Illumina and PacBio sequencing [4, 5, 11, 12]. This is frequently the case and we therefore strongly believe that the democratization of large-scale DNA barcoding will be important for upscaling species discovery across the globe and encouraging the use of the reverse workflow across many labs.

Oxford Nanopore's MinION has the potential to help with achieving these goals. It is a low-cost and portable real-time sequencing device. However, its use and reliability for large scale specimen handling remains to be fully understood. We recently showed that 500 reliable DNA barcodes can be obtained using 1D^2 sequencing on one flowcell of MinION. To our knowledge this was the highest number of products that were successfully multiplexed in a single MinION flowcell (5X higher than a study which examined ~100 amplicons of rDNA [11]) but such low throughput still meant that the cost per barcode remained high (ca. 2 USD: [13]). While this scale is useful for many species' identification projects, it is unlikely to be effective for large-scale species discovery where samples can contain thousands of specimens. Furthermore, the 500 specimens were barcoded using 1D^2 sequencing which requires a complicated library preparation, base-calling is computationally intensive, and application to amplicon sequencing is still under development. Unfortunately, it remained untested whether the more straightforward, but less accurate 1D sequencing can be used for large-scale species discovery. This is addressed in this manuscript.

However, we are here not only developing ways to use of MinION 1D sequencing for barcoding. Instead, we are also investigating the curious observation that only 466 species of phorid flies have been recorded for the Afrotropical Region [14]. Phoridae is a hyper-diverse clade belonging to the true flies (Diptera). Diptera is one of several hyper-diverse insect orders that also include beetles (Coleoptera), bees, wasps, and ants (Hymenoptera), and moths and butterflies (Lepidoptera). The species estimates for all of Insecta vary between 3 and 13 million (reviewed by Stork, 2018 [15]) with only ca. 1,000,000 currently being described [16]. Historically, the *inordinate fondness* of taxonomists for beetles has led to Coleoptera outpacing Diptera and Hymenoptera in numbers of described species. However, several recent studies suggest that Hymenoptera and Diptera are likely to be more species-rich. For example, Forbes et al. [17] hypothesize that Hymenoptera contained more species than either Diptera or Coleoptera based on parasite host ratios for Microhymenoptera, but this study showed an underappreciation for both the great numbers of Dipteran parasitoids and the diversity of true flies in general. Indeed, Diptera has recently been proven to be surprisingly rich in a number of large-scale biodiversity studies. In a large barcoding study of Canadian insects, Hebert et al. [18] found that Hymenoptera and Diptera together accounted for two thirds of the 46,937 BINS acquired and predicted that one dipteran family (Cecidomyiidae) has 16,000 species in Canada. The authors then extrapolated to the worldwide fauna which they estimated to be 10 million insect species, of which 1.8 million were predicted to be cecidomyiids [18]; i.e., a single family of Diptera may surpass the number of described species in all of Coleoptera. Other studies similarly hint at the extraordinary richness of Diptera. The Zurqui All Diptera Biodiversity Inventory (ZADBI) of a single site in Costa Rica was heavily reliant on specimens collected with two Malaise traps run for a one-year period [19]. Only

41,001 specimens (a small fraction of the hundreds of thousands collected) could be studied by taxonomic experts [20]. The specimens that were examined revealed 4,332 species of Diptera, of which 800 were from Cecidomyiidae and 404 were for Phoridae [20], the fly family of focus here.

Phoridae, or scuttle flies, are a worldwide family of true flies with approximately 4300 described species [14]. Over 1600 of these are in the giant genus *Megaselia* Rondani, which has been described as "one of the largest, most biologically diverse and taxonomically difficult genera in the entire animal kingdom" [21]. In groups like *Megaselia*, the species discovery problem appears insurmountable. Extremely large numbers of specimens are routinely collected, and they can belong to very large numbers of species. Even in urban and suburban habitats, the diversity of the family can be surprisingly high. Henry Disney, a world expert on the family, has recorded 75 species of phorids (48 of *Megaselia*) in his modest suburban garden [22]. Similarly, the BioSCAN project in Los Angeles found that backyards in the city supported as many as 82 species [22]. In natural areas, the diversity and abundance tend to be much higher and sorting such samples into species-level units using traditional workflows is very labor-intensive. Rare and new species are often hidden among very large numbers of common and described species. The study of groups like *Megaselia* requires examination of thousands of specimens for which prodigious notes have to be taken as specimens are compared. Many detailed drawings are prepared (for *Megaselia* drawings of male genitalia are essential) – often based on dissections and slide mounts – because many known/common species cannot be identified without detailed inspection. The process can take hours, days, or longer for a single species discovery. This traditional workflow thus often discourages all but the most tenacious taxonomists from taking up the study of hyper-diverse genera within insects.

Here, we test whether 1D MinION sequencing may offer a rapid and revolutionary approach to exploring phorid diversity more comprehensively. MinION sequencing is here applied to ca. 30% of the phorid specimens that were collected in a single Malaise trap in Kibale National Park, Uganda. We here describe how we processed ~8700 specimens, obtained ~7000 accurate barcodes, and found 650 species (of which almost 90% are *Megaselia*). All this could be accomplished using a workflow that requires less than a month.

RESULTS

MinION based DNA barcoding

The experiment was designed to obtain full-length COI barcodes for two sets of specimens via tagged amplicon sequencing. A total of 8699 specimens were processed (Set 1: 4275; Set 2: 4519; 95 specimens were shared between the sets) (Fig. 1). In order to assess amplification success rates, a subset of PCR products for each of the ninety-two 96-well plates were assessed with agarose gels. The extrapolated amplification success rates were 86% and 74% for the two sets of specimens (80.7% overall); i.e., we estimated that >3600 and >3300 DNA barcodes should be obtainable via MinION sequencing given that gels tend to underestimate amplification success rates (Table 1). The PCR products for each set were pooled and sequenced using MinION (set 1: 7,035,075; set 2: 7,179,121 1D nanopore reads). Both sets were sequenced in two sequencing runs. The first run for each set was based on the pooled PCR products for all specimens in the sets. It generated 3,069,048 and 4,853,363 reads, respectively. The results of the first run were used to estimate coverage for each PCR product. Products with weak coverage (<=50x) were re-pooled and re-sequenced (set 1: 2172 amplicons; set 2: 2211 amplicons). This added 3,966,027 and 2,325,758 reads to each set and improved the coverage of many low-coverage barcodes (Fig. 2). The combined data were processed using an improved version of the bioinformatics pipeline in Srivathsan et al. [13]. The improvements led to a higher demultiplexing rate (14%

increase for set 1: 898,979 vs. 787,239 reads; 9% increase for set 2: 647,152 vs. 593,131 reads) and faster demultiplexing (10X using 4 cores: demultiplexing in 9 min vs 87 min for one of the datasets).

Demultiplexing of all data and preliminary barcode calling revealed 3,797 and 3,476 MAFFT barcodes with >=5X coverage and <1% ambiguous bases. These barcodes were subject to correction using RACON [23] which yielded the same number of barcodes. When the barcodes for the two sets of samples were combined, we overall obtained 7,220 MAFFT and RACON barcodes. These preliminary barcodes still contain indel and substitution errors that were addressed with an amino-acid correction pipeline that was first implemented in Srivathsan et al. [13]. It yielded 7,178 MAFFT+AA and 7,194 RACON+AA barcodes. Some barcodes were not retained because this pipeline rejects barcodes that have five or more consecutive indel errors. Finally, the two sets of corrected barcodes were consolidated. This yielded a set of 7,115 consolidated barcodes. We rejected barcodes where the alignment of MAFFT+AA and RACON+AA barcodes required the insertion of indels, as the +AA barcodes are expected to be indel-free. Such indels in the alignments of MAFFT+AA and RACON+AA barcodes also indicate discrepancies between MAFFT and RACON barcode estimates and there is no objective reason to prefer one barcode over the other. The overall barcoding success rate was thus 81.9% (7,115 barcodes for 8,699 specimens). This was close to the expected 80.7% success rate based on gel electrophoresis; i.e., MinION sequencing consistently produced sequence data for successfully amplified products. A subsequent contamination check via BLAST revealed that of the 7,115 barcodes, 53 barcodes were unlikely to pertain to phorid flies (<1%) and we thus retained 7,062 barcodes for species richness estimation. Lastly, we inspected the reads obtained for the 92 negative controls (1 per microplate). Five negatives yielded MAFFT barcodes. Four of these had a >97% match to non-phorids (two humans, one fish, one mollusc) and were eliminated. One low coverage (13X) negative survived all filters and matched phorid COI. It was removed after ascertaining that it did not impact the accuracy of the barcodes in the plate. This could be tested by comparing the MinION barcodes for the plate with Illumina barcodes obtained from different PCR products for the same DNA extraction plate (see below).

Accuracy and selection of barcode sets

To find the best strategy for obtaining accurate barcodes, we compared 5 sets of barcodes (MAFFT, RACON, MAFFT+AA, RACON+AA, and consolidated barcodes) with the corresponding barcodes based on Illumina sequencing. Illumina barcodes were obtained for 6,373 specimens for the same specimens using different primers that amplified a 313 bp subset of the full-length barcodes. The comparisons showed that the uncorrected MAFFT and RACON barcodes had an accuracy of 99.61% and 99.51% (Table 2). Correction of these barcodes using the amino-acid correction pipeline improved the accuracy considerably (>99.9% in all cases). The barcodes were corrected after testing several “namino” parameters. Overall, namino=2 was found to yield the most accurate barcodes, and minimized the number of inaccurate barcodes. We found that MAFFT+AA barcodes were more accurate than RACON+AA barcodes, but MAFFT+AA barcodes contained a much higher number of ambiguous nucleotides (Fig. 3). When RACON+AA and MAFFT+AA barcodes were consolidated, the resulting “consolidated barcodes” were found to be highly accurate (99.99%) and contained few ambiguous bases (median = 0.3 %).

We furthermore compared the mOTU richness estimated by the different barcode sets. mOTU richness was very similar across MinION and Illumina barcodes for the consolidated and uncorrected MAFFT/RACON barcodes. MAFFT+AA barcodes performed well in this comparison, but yielded fewer mOTUs than Illumina barcodes: this may be due to a higher proportion of ambiguous nucleotides (Fig. 3). However, comparison of mOTU richness alone does not imply the same specimens were grouped into mOTUs across MinION and Illumina barcode sets. We thus also calculated the match ratio for the datasets (3% clustering threshold). We found that all five barcode sets (MAFFT, RACON, MAFFT+AA, RACON+AA, and consolidated barcodes, namino=2) also had high match ratios (>0.97) with the consolidated barcodes and RACON barcodes performing best with match ratios of >0.98 (consolidated barcodes: 0.989, RACON: 0.989, RACON+AA: 0.982). However, upon closer inspection only the multiple sequence alignment (MSA) of the consolidated barcodes was indel-free while the other MSAs contained indels. The largest number of indels was found in the MSA of uncorrected RACON barcodes which indicates that the RACON barcodes retain a fair number of indels and may not be of sufficient quality for submission to sequence databases. Overall, we would thus recommend the usage of consolidated barcodes as the final barcode set. Because they maximize the per-base accuracy, estimated mOTU diversity, match ratios, and yield high-quality alignments.

Species richness estimation

We thus proceeded to characterize the diversity of the phorid flies collected from the Malaise traps based on the consolidated barcodes (namino=2). We overall obtained a mean of 683 mOTUs (2%: 728, 3%: 685, 4%: 636) when the thresholds were varied from 2-4%. Species accumulation and Chao 1 curves for mOTUs at 3% were not found to have reached a plateau, but the shape of the curves suggest an estimated diversity is >1000 species in a single site collected by one Malaise trap (Fig. 4).

Congruence with morphology

mOTUs are expected to be affected by mistakes caused by lab contamination and species delimitation errors due to the biological properties of barcodes. We find that 6 of the 100 clusters contained a single misplaced specimen and that there was one small cluster of four specimens that appeared to consist of a mixture three morpho-species. This implies that 9 of the >1500 examined barcoded specimens were misplaced due to lab contamination. mOTUs based on barcodes are expected to underestimate species for those that recently speciated and overestimate species with deep splits [24]. This means that taxonomists working with mOTUs should check for signs of lumping and splitting for closely related taxa. Our initial morphological check of 100 randomly selected clusters (>1500 specimens) took ca. 30 hours. This preliminary morphological screening covered all specimens while future studies could concentrate on complex clusters, as any differences that may have occurred within 5% clusters were not discernible without further preparations (dissection and slide mounting). Identifying these areas of potential ambiguity is an advantage of this pipeline. It allows taxonomic experts to focus time and energy on these complex clusters.

New Species Description

A primary aim of the reverse workflow is finding rare new species for description in bulk samples. With specimens pre-sorted into mOTUs, morphologists have easy access to interesting and potentially rare species that would otherwise remain undiscovered. While examining the 100 mOTUs, eight specimens were found to belong to a distinctive new species of *Megaselia*. This species provided a good opportunity to demonstrate how the reverse workflow aids in species discovery because a mOTU-specific haplotype network informed on which specimens should be studied with morphology. The species is here described, and the description incorporates the molecular data.

Megaselia sepsioides Hartop sp. n.

urn:lsid:zoobank.org:pub:ED268DF2-A886-4C31-A4FB-6271C382DECE

Description

See Fig. 5, 6.

In an effort to continue reducing redundancy and ambiguity in species descriptions, the description of this species has excluded the character table from the method previously established for *Megaselia* [25-27] and uses a barcode and photographic description. Photographs are a key element in descriptions for large, diverse groups [28], where verbose descriptions can be both overwhelmingly time consuming and insufficiently diagnostic. Most characters that would have been in table form are clearly visible in the photographs provided; the few that are not were deemed irrelevant to either the description or diagnosis of the species.

Diagnosis

This species is unmistakable even within the gargantuan and taxonomically difficult genus *Megaselia*. The semi-circular expansion with modified peg-like setae on the forefemur is unique among described members of the genus (Fig. 5, b). Similarly, the severe constriction of the hind tibia basally is diagnostic (Fig. 5, d and e). The narrow and elongate form of the abdomen is notable. Fig. 6 shows variations in setation between haplotypes. Only single specimens of the two distinct haplotypes are available; more specimens will be necessary to determine if these are eventually removed as distinct species or fall within a continuum of intraspecific variation.

DNA barcode for UGC0005996:

actttatattttatttttggagcttgagctggaatagtaggtacttccttaagaatcataattcgtgctgaattaggacacccaggagcacttattggtgatgaccaaatttataatgtgattgttactgcacatgcttttattataattttttttatagtaatacctattataataggaggttttggtaattgacttgtacctttaatattaggagccccagatatggcattccctcgaatgaataatataagtttttgaatattacctccttctttaactcttttattagccagaagtatagtagaaaatggagctggaactggttgaacagtttatcctcctttatcttctagaatcgctcatagtggagcttctgttgatttagcaattttctctcttcatttagctggaatttcatctattttaggagctgtaaattttattacaacaattattaatatacgatcatcaggtattacatttgaccgaatacctctatttgtttgatctgtaggtattacagctttattgctactcttatcacttcctgttttagctggtgctattacaatactattaacagaccgaaattttaatacttcattttttgacccagcaggaggaggagatccaattttataccaacatttattc

Material Examined

Holotype:

Scientific Name: *Megaselia sepsioides* Hartop 2019; country: Uganda; state Province: Kamwenge; locality: Kibale National Park: 1530 m; Coordinates: 00°33'54.2"N 30°21'31.3"E; sampling protocol: Malaise trap; event date: iii-xii, 2010; individual count: 1; sex: male; life stage: adult: UGC0005996; identified by: Emily Hartop: 2019; institution code: LKCNHM; collection code: UGC; basis of record: preserved specimen.

Paratypes:

Scientific name: *Megaselia sepsioides* Hartop 2019; country: Uganda; state province: Kamwenge; locality: Kibale National Park; elevation: 1530 m; coordinates: 00°33'54.2"N 30°21'31.3"E; sampling protocol: Malaise trap; event date: iii-xii, 2010; individual count: 7; sex: male; life stage: adult; catalog number: UGC0012899, UGC0012244, UGC0012568, UGC0003003, UGC0005864, UGC0012937, UGC0012971; identified by: Emily Hartop; date identified: 2019; institution code: LKCNHM; collection code: UGC; basis of record: preserved specimen

Distribution

Known from a single site in Kibale National Park, Uganda.

Biology

Unknown.

Etymology

Named by Yuchen Ang for the sepsid-like (Diptera: Sepsidae) foreleg modification.

DISCUSSION

Large scale species discovery using MinION

Our results suggest that MinION's 1D sequencing yields data of sufficient quality for producing high-quality DNA barcodes that can be used for large-scale species discovery. Through the development of new primer-tags, sequencing strategies, and improved bioinformatics procedures, we here increase the barcoding capacity of a MinION flowcell from 500 specimens (Srivathsan et al., 2018: $1D^2$ sequencing) to ~3500 specimens. This is achieved without a drop in accuracy because the new error correction pipeline is effective at eliminating most of the errors in the 1D reads (ca. 10%). Indeed, even the initial MinION barcodes (MAFFT & RACON) have very high accuracy (>99.5%) when compared to Illumina data. Note that this accuracy is even higher than what was obtained with $1D^2$ sequencing in Srivathsan et al. (2018: 99.2%). We suspect that this partially due to improvements in MinION sequencing chemistry and base-calling, but our upgraded bioinformatics pipeline also helps because it increases coverage for the amplicons. These findings are welcome news because 1D library preparations are much simpler than the library preps for $1D^2$. In addition, $1D^2$ reads are currently less suitable for amplicon sequencing (https://store.nanoporetech.com/kits-250/1d-sequencing-kit.html) [29].

We tested a range of different techniques for obtaining barcodes from 1D reads. Some techniques are computationally more expensive than others which raises the question of which bioinformatics pipeline should be used for obtaining accurate barcodes. Based on the current performance of MinION, we would recommend the usage of the "consolidated barcodes". They contain no indel errors and fewer substitution errors (99.99% accuracy) when compared to MAFFT+AA and RACON+AA barcodes. The mOTUs delimited with consolidated barcodes are virtually identical with the ones obtained via Illumina sequencing (Number of 3% mOTUs for Illumina: 661; MinION:

658; match-ratio: 0.989). In addition, consolidated barcodes were also the most accurate barcodes based on 1D^2 reads [13]. Nevertheless, we found that all barcode sets gave reliable mOTU estimates. For corrected barcodes (MAFFT+AA and RACON+AA) this was consistent with >99.9% per base accuracy, but the higher error rates of the uncorrected barcodes rarely affected mOTU estimates. This is because indels were treated as missing data and mOTU estimation only requires accurate estimates of distances for closely related taxa. This explains why the 99.5% accurate RACON barcodes can group specimens into mOTUs in a very similar manner to what is obtained with Illumina barcodes (match ratio of >0.98). The results imply that accurate mOTU estimates can be obtained even based on preliminary barcodes.

Based on the procedures described here, MinION barcodes can be generated rapidly and at a low sequencing cost of <0.35 USD per barcode. These properties make MinION a valuable tool for species discovery whenever a few thousand specimens have to be sorted to species (<5000). Even larger-scale barcoding is probably still best tackled with Illumina short-read or PacBio's Sequel sequencing [4, 5, 7] because the cost is lower and the quality of the reads is higher. However, both require access to expensive sequencers, the sequencing has to be outsourced, and the users usually have to wait for several weeks in order to obtain the data. This is not the case for barcoding with MinION where most of the data are collected within 10 hours of sequencing. Our proposed MinION pipeline has the additional advantage that it only requires basic molecular lab equipment including thermocyclers, magnetic rack, Qubit and potentially a server computer. The latter is only needed for base-calling and should be replaceable by a new portable, custom-built computational device for base-calling ONT data ("MinIT") while demultiplexing and barcode-calling only requires a regular laptop computer. Overall, these requirements are minimal which means that fully functional barcoding labs can be established for <USD 5,000. Arguably, the biggest operational issue may be access to a sufficiently large number

of thermocyclers given that a study of the scale described here involved amplifying PCR products in 92 microplates (=92 PCR runs).

Our new workflow for large-scale species discovery is based on sequencing the amplicons in two sequencing runs. The second sequencing run can re-use the flowcell that was used for the first run. Two runs are desirable because it improves overall barcoding success rates. The first run is used to identify those PCR products with “weak” signal (=low coverage). These weak products can then be re-sequenced in the second run. This dual-run strategy is designed to overcome the challenges related to sequencing large numbers of PCR products: the quality and quantity of DNA extracts are poorly controlled and PCR efficiency varies considerably. Pooling of products ideally requires normalization, but this is not practical when thousands of samples are handled. Instead, one can use the real-time sequencing provided by MinION to determine coverage and then boost the coverage of low-coverage products by preparing and re-sequencing a separate library that contains only the low coverage samples. Given that library preparations only require <200 ng of DNA, even a set of weak amplicons will yield a sufficient amount of DNA.

MinION sequencing and the “reverse workflow”

Based on our data, we would argue that MinION is now suitable for implementing the “reverse-workflow” where all specimens are sequenced first before mOTUs are assessed for morphological consistency. This differs from the traditional workflow in that the latter requires sorting based on morphology with only some morpho-species being subsequently tested via limited barcoding. We would argue that the reverse-workflow is more suitable for handling species- and specimen-rich clades because it requires less time than high-quality sorting based on morphology which often involves genitalia preparations and slide-mounts. We would argue that expert sorting and identification of material based on morphology is only very efficient for taxa where species-specific morphological characters are easily accessible and the number of specimens is small. However, most undiscovered and undescribed species are in poorly explored groups and regions where specimen sorting and identification using the traditional techniques will be much slower. For example, even if we assume that an expert can sort and identify 50 specimens of unknown phorids per day, the reverse workflow pipeline would increase the species-level sorting rate by >10 times (based on the extraction and PCR of six microplates per day). In addition, the sorting would be carried out by lab personnel trained in amplicon sequencing while accurate morpho-species sorting requires highly specialized taxonomic experts. Even such highly trained taxonomic experts are often not able to match males and females of the same species (often one sex is ignored in morphological sorting) while the matching of sexes and immatures is an automatic and desirable byproduct of applying the reverse workflow [6].

One key element of the reverse workflow is that vials with specimens that have haplotype distances <5% are physically kept together. This helps when taxonomic experts scrutinize mOTUs for congruence with morpho-species. Indeed, graphical representation of haplotype relationships (e.g., haplotype networks) can be used to guide morphological re-examination as illustrated in our description *of Megaselia sepsioides* (Fig. 7). The eight specimens belonged to seven haplotypes, the most distant haplotypes were dissected in order to test whether the data are consistent with one or two species. Variations in setation were observed (Fig. 6) that were deemed likely to be consistent with intraspecific variation. Note that the morphological examination of clusters is straightforward because the use of QuickExtract for DNA extractions ensures morphologically intact specimens. We predict that taxa that have been historically ignored or understudied due to extreme abundance of common species and high species diversity will now become more accessible because taxonomic experts can work on pre-sorted material instead of thousands of unsorted specimens.

But how good is the correspondence between mOTUs and morpho-species. We checked 100 randomly chosen mOTUs for congruence with morphospecies. We find that 93% of clusters are congruent, and over 99% of specimens (six of the seven cases of incongruence were single specimens). This is in line with congruence levels that we observed for ants and odonates [4, 6]. This means that MinION barcodes are not only reliable but also yield mOTUs that are largely congruent with morphospecies. This also means they are suitable for characterizing even complex and species-rich samples of hyper-diverse invertebrate clades.

Remarkably high diversity of Phoridae in Kibale National Park

It is astonishing that the barcodes obtained from a single site in Kibale National Park (Uganda) revealed ca. 650 mOTUs of phorid flies. This diversity, obtained from one Malaise trap, contains 150% of the number of described phorid species (466) known from the entire Afrotropical region; i.e., numerically at least 184 species must be new to the region or to science [14]. Note that the barcoded specimens only represent 8 one-week samples between March 2010 and February 2011; i.e., forty-four samples obtained from the same Malaise trap remain un-sequenced. We thus expect the diversity from this single site to eventually well exceed 1000 species (Fig. 4).

The unexpectedly high species richness found in this study inspired us to muse about the species diversity of phorids in the Afrotropical region. This is what Terry Erwin did when he famously estimated a tropical arthropod fauna of 30 million species based on his explorations of beetle diversity in Panama [30]. Such speculation is useful because it raises new questions, inspires follow-up research, and may be needed given that even with extensive sampling, the species richness of diverse taxa can be remarkably difficult to estimate [31]. The Afrotropical region comprises roughly 2000 squares of 100 km^2 size. In our study we only sampled a tiny area within one of these squares and observed 650 species which are likely to represent a species assemblage that exceeds >1000 species. This will only be a subset of the phorid fauna in the area because many specialist species (e.g., termite inquilines) are not collected in Malaise traps. Of course, the 1000 species are also only a subset of the species occurring in the remaining habitats in the same 100 km^2 square which is likely to be home to several thousand species of phorids. But let's only assume that on average each of the two-thousand 100 km^2 squares have 100 endemic phorid species. If so, the "endemic" phorids alone would contribute 200,000 species of phorids to the Afrotropical fauna without even considering the contributions to species diversity by the "normal" species turnover of the remaining species. These considerations make us believe that it would be somewhat surprising if the Afrotropical region were to have fewer than half a million species of phorids! Based on our sample from Kibale National Park, 90% of species and specimens would belong to *Megaselia* as currently circumscribed. Could it be that there are 450,000 species of Afrotropical *Megaselia*? These guestimates would only be much lower if the vast majority of phorid species had very wide distributions which we consider somewhat unlikely given that the Afrotropical region covers a wide variety of climates and habitats.

As documented by our study, the bulk of phorid species are *Megaselia*. Based on the traditional workflow, almost all these specimens would be relegated to unsorted Malaise trap residues for decades or centuries. Indeed, there are thousands of vials labeled as "Phoridae" decorating the shelves of all major museums worldwide. The traditional workflow is not able to keep pace with the species numbers and abundance. This makes rapid species-level sorting with "NGS barcodes" [4] so important. Biologists will finally be able to work on taxa that are so specimen- and species-rich that they were considered unworkable with the traditional techniques. MinION barcodes will be one of the techniques that can be used to tackle these clades. We predict that these barcodes will become particularly important for setting up mobile laboratories that can operate under field conditions, but MinION barcodes can also be generated by highschool students and citizen scientists.

METHODS

1. Sampling

Samples were collected from a single Townes-type Malaise trap [32], in the Kibale National Park, close to Kanyawara Biological Station in the evergreen primeval forest at an altitude of 1513 m (00°33'54.2"N 30°21'31.3"E) (Fig. 4). Kibale National Park is characterized as a fragment of submontane equatorial woodland being home to 216 tree species [33]. Temperatures in Kibale range from 16°C to 23°C (annual mean daily minimum and maximum, respectively) [34]. As described, the Malaise trap was checked every week when the collecting bottle with the material was replaced by a resident parataxonomist ([35]: Mr. Swaibu Katusabe). The material in Kibale National Park in Uganda was collected and transferred in accordance with approvals from the Uganda Wildlife Authority (UWA/FOD/33/02) and Uganda National Council for Science and Technology (NS 290/ September 8, 2011), respectively. The material was thereafter sorted to higher-level taxa. Target Diptera taxa were sorted to family and we here used Phoridae. The sampling was done over several months between 2010 and 2011. For the study carried out here, we only barcoded ca. 30% of the phorid specimens. The flies were stored in ethyl alcohol at -20–25°C until extraction.

2. DNA extraction

DNA was extracted using 10 ul of QuickExtract in a 96 well plate format and the whole fly was used to extract DNA. The solution with the fly was incubated at 65°C for 15 min followed by 98°C for 2 min. No homogenization was carried to ensure that intact specimen was available for morphological examination.

3. MinION based DNA barcoding

I. Polymerase Chain Reactions (PCRs)

Each plate with 96 QuickExtract extracts (95 specimens and 1 control, with exception of one plate with no negative and one partial plate) was subjected to PCR in order to amplify the 658 bp fragment of COI using LCO1490 5' GGTCAACAAATCATAAAGATATTGG 3' and HCO2198 5' TAAACTTCAGGGTGACCAAAAAATCA 3' [36]. Each PCR product was amplified using primers that included a 13 bp tag. For this study, 96 thirteen-bp tags were newly generated in order to allow for upscaling of barcoding; these tags allow for multiplexing >9200 products in a single flowcell of MinION through unique tag combinations (96x96 combinations). To obtain these 96 tags, we first generated one thousand tags that differed by at least 6 bp using with BarcodeGenerator (http://comailab.genomecenter.ucdavis.edu/index.php/Barcode_generator) [37] However, tag distances of >6 bp are not sufficient because they do not take into account MinION's propensity for generating homopolymer and other indel errors. We thus excluded tags with homopolymeric stretches that were >2 bp long. We next used a custom script that identified tags that differed from each other by indel errors and eliminated tags recursively to ensure that the final sets of tags differed from each other by >=3bp errors of any type (any combination of insertions/deletions/substitutions). This corresponded with our bioinformatic strategy of using <=2bp mismatch in tag identification. Lastly, we excluded tags that ended with "GG" because LCO1490 starts with this motif. The PCR conditions were as follows, reaction mix: 10 µl Mastermix from CWBio, 0.16 µl of 25mM $MgCl_2$, 2 µl of 1 mg/ml BSA, 1 µl each of 10 µM primers, and 1ul of DNA. The PCR conditions were 5 min initial denaturation at 94°C followed by 35 cycles of denaturation at 94°C (30 sec), 45°C (1 min), 72°C (1 min), followed by final extension of 72°C (5min). For each plate, a subset of 7-12 products were run on a 2% agarose gel to ensure that PCRs were successful. Of the 96 plates studied, 4 plates were excluded from further analyses as they had <50% amplification success and one plate was duplicated across the two runs by accident.

II. MinION sequencing

The most cost-effective strategy for nanopore sequencing was optimized during the study. For the initial experiment (set 1) we sequenced amplicons for 4275 phorid flies. The flowcell was used for 48 hours and yielded barcodes for ~3200 products, but we noticed lack of data for products for which amplification bands could be observed on the agarose gel. We thus re-pooled all products with a sequencing depth <=50X (2119 specimens), prepared a new library and sequenced them on a new flowcell. The experiment was successful. However, in order reduce sequencing cost, we pursued a different strategy for the second set of specimens. This set consisted of pooled amplicons for 4519 flies, but here we stopped the sequencing on the flowcell after 24 hours. The flowcell was then washed using ONT's flowcell wash kit and prepared for reuse. The results from the first 24 hours of sequencing were used to identify amplicons with weak coverage. They were re-pooled, a second library was prepared, and sequenced on the pre-used and washed flowcell.

Amplicon pooling strategy: For set 1, all plates were grouped by amplicon strength as judged by the intensity of products on agarose gels (5 strong pools +2 weak pools). For set 2, each plate was pooled, quantified, and cleaned using either 1X Ampure beads or 1.1X Sera-Mag beads in PEG. For the re-sequencing of weak or "problematic" products (see below), we identified the latter based on the results of the initial sequencing run. We located (1) specimens <=10X coverage (set 1: 1054, set 2: 1054) and (2) samples with coverage between 10X and 50X (set 1: 1118, set 2: 1065). Lastly, we also created a (3) third pool of specimens with problematic products that were defined as those that were found to be of low accuracy during comparisons with Illumina barcodes and those that had high levels of ambiguous bases (>1% ambiguous bases during preliminary barcode calling). Very few amplicons belonged to this category (set 1: 68, set 2: 92) and it is thus not included in the flowchart in Figure 1. In order to efficiently re-pool hundreds of specimens across plates we wrote a script that generates visual maps of the all microplates that illustrate where the weak products are found.

Library preparation and sequencing: We used the SQK-LSK109 ligation sequencing kit for library preparation and sequencing. Our first experiment on set 1 used 1 ug of starting DNA while all other libraries used 200 ng pooled product. Library preparation was carried out as per manufacturer's instructions with one exception: the various clean-up procedures at the end-prep and ligation stages used 1X Ampure beads instead of 0.4 X as suggested in the instructions because the amplicons in our experiments were short (~735 bp with primers and tags). The sequencing was carried out using MinION sequencer with varying MinKNOW versions between August 2018 - January 2019. Fast5 files generated were uploaded onto a computer server and base-calling was carried out using Guppy 2.3.5+53a111f. No quality filtering criteria were used. Our initial work with Albacore suggested that quality filtering improved demultiplexing rate but overall more reads could be demultiplexed without the filtering criterion.

III. Data analyses for MinION barcoding

We analysed the data using miniBarcoder [13], which was improved in several ways for the present study. Overall, the pipeline starts with a primer search with *glsearch36,* then flanking nucleotide sequences are identified, and reads are demultiplexed based on tags. For the latter, an error of up to 2 bp errors are allowed. The demultiplexed reads are aligned using MAFFT v7 (--op 0) (here v7) [38]. In order to improve speed, we used only a random subset of 100 reads from each demultiplexed file for alignment. Based on these alignments, a majority rule consensus is called to obtain what we call “MAFFT barcodes”. A second preliminary barcode is generated by mapping all reads back to the MAFFT barcode using Graphmap (here v0.5.2) [39] and calling the consensus using Racon (here, v1.3.1) [40]. This yields what we call “RACON barcodes”. Both MAFFT and RACON barcodes are subject to further correction based on publicly available barcodes in GenBank. These corrections are advisable in order to fix remaining indel errors. The correction takes advantage of the fact that COI sequences are translatable; i.e., an amino-acid based error correction pipeline can be used (details can be found in Srivathsan et al. (2018). Applying this pipeline to MAFFT and RACON barcodes respectively yields MAFFT+AA and RACON+AA barcodes. Lastly, these barcodes can be consolidated by comparing them to yield “consolidated barcodes”.

The version of the pipeline in Srivathsan et al. [13] was modified as follows:

a. *Tackling 1D reads for massively multiplexed data:* The large number of simultaneously barcoded specimens presented many challenges related to varying coverage and quality. We hence sought to develop ways to account for increased error rates of 1D sequencing and develop objective ways for quality assessments based on the MinION data and publically available data (GenBank): (1) The GraphMap max error was increased from 0.05 to 0.15 to account for error rates of 1D reads. (2) We modified the approach for calculating consolidated barcodes. We here use the strict consensus of MAFFT+AA and RACON+AA barcodes in order to resolve conflicts between MAFFT and RACON barcodes if there are substitution conflicts. In Srivathsan et al. (2018) we accepted MAFFT+AA barcodes in cases of conflict, but for the 1D data we found that MAFFT+AA barcodes had more ambiguities than RACON+AA barcodes which could be resolved via calculating a strict consensus. We also introduced a criterion for identifying “problematic” barcodes based on observing too many differences between MAFFT+AA and RACON+AA barcodes; i.e., if indels are found in the alignment of MAFFT+AA and RACON+AA barcode during consolidation, the barcode is rejected. This criterion is based on the fact that +AA barcodes should be indel-free. Either the MAFFT+AA or the RACON+AA barcode is likely to be incorrect, but there is no objective way to identify the correct sequence. (3) We assessed how different window sizes can impact the amino-acid correction pipeline by varying the “namino” parameter (number of AA to be inspected in each side of an indel).

b. Demultiplexing rate: (1) We introduced a “homopolymer compression” of the putative tag sequences in order to improve demultiplexing rates. After primer searches, the old pipeline used to identify the flanking 13 bps that were likely to be tag sequence. Instead we now use a 20 bp flanking sequence and then compress any homopolymer >3bp before searching for 13 bp tags. (2) We now analyze reads that are likely the result of two products that ligated during library preparation. This was based on our experience with one library where ligation of products was prominent. Long reads are split into size classes of <1300, >1300 and <2000. These settings were set based on 658 bp barcode of COI: the total product size including tags and primers is 735 bp, and hence, a sequence with two products ligated to each other is expected to be 1470 bp long. The sequences were split in a manner that ensured that the tags of first product are not affecting the tag found in the second, i.e., primer search for second product was conducted after the first 650 bp of the sequence. Currently, this option is only available for reads that consist of two ligated products.

c. Processing speed and memory requirements: (1) For primer identification we now limit the searches to the first and last 100 bp of each read which allowed for improving speed of the primer search. (2) We parallelized the process of demultiplexing and MAFFT barcode calling using multiprocessing in python. This allows for fast demultiplexing whenever computational power is limited. (3) We optimized the pipeline with regard to memory consumption and ensured that all processes are applied to batches of <=20000 sequences. The latter modification is needed given the rapid increase in MinION output and batch processing is scalable with increased output.

4. Illumina Based NGS Barcoding for validation

In order to validate MinION barcodes and optimizing error correction strategies, we used reference COI barcodes obtained via Illumina sequencing for the same QuickExtract DNA extractions. Illumina sequencing was carried out for a 313 bp fragment of the same COI barcoding region using m1COlintF: 5'-GGWACWGGWTGAACWGTWTAYCCYCC-3' [41] and modified jgHCO2198: 50-TANACYTCNGGRTGNCCRAARAA YCA-3 [42]. We recently conducted an extensive analyses to understand if a 313 bp minibarcode is able to provide similar identifications and species delimitations as the 658 bp barcodes and found this to be the case when examining >5000 species [43]. Tagged primers were used for the PCRs as specified in Wang et al. (2018) [44]. The PCR mix was as follows: 4 µl Mastermix from CWBio, 1µl of 1 mg/ml BSA, 1µl of 10 µM of each primer, and 1ul of DNA. PCR conditions: 5 min initial denaturation at 94°C followed by 35 cycles of denaturation at 94°C (1 min), 47°C (2 min), 72°C (1 min), followed by final extension of 72°C (5 min). The PCR products were pooled and sequenced along with thousands of other specimens in a lane of HiSeq 2500 (250 bp PE sequencing). The data processing followed Wang et al. (2018): paired end reads were merged using PEAR (v 0.9.6) [45], reads were demultiplexed by an inhouse pipeline that looks for perfect tags while allowing for 2 bp mismatches in primer sequence. For each sample, demultiplexed reads are merged to identify the most dominant sequence, and a barcode is accepted only if this sequence is 5X as common as the next most common sequence.

5. Assessment of MinION barcodes and mOTU delimitations

Both MinION and Illumina barcodes were subject to contamination check. For MinION barcodes we used preliminary MAFFT barcodes given that this is the largest barcode set. Barcodes were matched to GenBank using BLASTN and taxonomic classifications were assigned using readsidentifier [46]. Any barcode with >95% match to a non-phorid sequence was excluded from the dataset. Furthermore, if any barcode best matched to Bacteria, it was also excluded.

MinION barcodes were assessed by scripts provided in the miniBarcoder package (assess_corrected_barcodes.py and assess_uncorrected_barcodes.py). For uncorrected barcodes, this was done by aligning barcodes to reference Illumina barcodes using dnadiff [47]. For corrected barcodes (+AA), we used MAFFT to obtain pairwise alignments. This allowed us to compare per base accuracy. Here we excluded those barcodes that had >3% difference between MinION and Illumina barcodes. These are likely due to wet lab errors (<0.5% of specimens). Such error is not entirely surprising given that the MinION and Illumina barcodes were generated using different amplicons.

We were further interested in understanding how mOTU delimitation is affected by error correction. Barcodes were aligned using MAFFT and MinION barcodes were further trimmed to the 313 bp region of the Illumina barcode. mOTU delimitation was done at 2, 3,and 4% using SpeciesIdentifier (objective clustering) [48]. mOTU richness was estimated and we furthermore calculated match ratio between two sets of clusters [49]. Match ratio is given by $\frac{2N_{match}}{N_1+N_2}$.

6. Morphological examination

For morphological examination of the clustered specimens we used 100 randomly selected non-singleton mOTUs delimited at 5% but also kept track of sub-clusters within the large mOTUs that were splitting at 1-5%. This allowed for examination of closely related, but potentially distinct, species. We were mostly interested in understanding if MinION barcodes were placing specimens into mOTUs incorrectly and hence we examined if specimens were consistent morphologically in each of these 5% clusters. The choice of 5% cluster may seem initially inconsistent with the choice of 3% for mOTU delimitation for other analyses, but examination of all specimens within 5% clusters allows for comparing multiple closely related 3% (or 1-5%) mOTUs. This often requires genitalia preparations which will be carried out at a larger scale once more specimens have been sequenced. In this study, the process is only illustrated for the newly described species for which we illustrate how the haplotype network obtained with the median joining method in PopART (Fig. 7) [50] guides morphological examination of specimens for the new species, *Megaselia sepsioides* sp. nov.. Specimens with conspicuously large haplotype differences were dissected in order to rule out the presence of multiple closely related species. Differences in setation were observed between the two distant haplotypes (UGC0012899 and UGC001224) and the main cluster (UGC0003003, UGC0005864, UGC0005996, UGC0006224, UGC0012568, UGC0012937, and UGC0012971) and are detailed in Fig. 6. It is deemed likely that with further collection of specimens from this haplotype cluster, these now rather distinct differences will be absorbed into a continuum of intraspecific variation. Specimen examination was done with a Leica m80 and Leica M205 C stereo microscopes and the images were obtained with a Dun Inc. Microscope Macrophotography system (Canon 7D chassis with 10X Mitutoyo lens). Photography stacking was done with Zerene stacker. Specimens were cleared with clove oil and mounted on slides in Canada balsam following the protocol of Disney [51].

7. Species Richness estimation

The most accurate set of barcodes were used for assessing the overall diversity of the barcoded phorids. mOTU delimitation was based on SpeciesIdentifier [48]. The 3% mOTUs were used for estimating species richness. Here we used EstimateS9 [52], and changed the diversity settings to use the classical formula of Chao1 and Chao2. This is because coefficient of variation of the abundance or incidence distribution was >0.5.

AVAILABILITY OF SOURCE CODE AND REQUIREMENTS

- Project name: miniBarcoder
- Project home page: https://github.com/asrivathsan/miniBarcoder

- Operating system(s): Linux, MacOSX (for initial barcode calling)
- Programming language: Python
- Other requirements: MAFFT BARCODE: MAFFT, glsearch36, RACON BARCODE: GraphMap, Racon, and amino acid correction: BioPython, MAFFT
- License: GNU GPL
- New scripts included: mb_parallel_consensus.py mb_parallel_demultiplex.py, repool_by_plate.py
- Updated scripts: consolidate.py, bug fix in aacorrection.py

AVAILABILITY OF SUPPORTING DATA AND MATERIALS

Besides the source code, the nanopore data and input files and barcode sets will be made available once FTP link is obtained.

DECLARATIONS

Competing Interests: The author(s) declare(s) that they have no competing interests.

Funding: We would like to acknowledge support from the following grants: MOE grant for biodiversity discovery (R-154-000-A22-112), NUS SEABIG grant (R-154-000-648-646 and R-154-000-648-733), and institutional research funding (IUT21-1) of the Estonian Ministry of Education and Research for O.K..

Author's Contributions: R.M. and A.S. conceived the workflow and the analytical approach. O.K. conducted/organized the sampling and Diptera sorting and commented on the manuscript. Molecular work was conducted by J.P., W.T.L., S.N.K., E.H..and A.S. Pipeline development and data analyses was conducted by A.S. Morphological examination, and species description was conducted by E.H. Figures were prepared by E.H., S.N.K., A.S. Manuscript was written by R.M., A.S. and E.H..

Acknowledgements: The material in Kibale National Park in Uganda was collected and transferred in accordance with approvals from the Uganda Wildlife Authority (UWA/FOD/33/02) and Uganda National Council for Science and Technology (NS 290/ September 8, 2011), respectively. Yuchen Ang is thanked for his help photographing the new species and for his suggestion of sepsioides as the specific epithet. Henry Disney and Brian Brown are thanked for their consultation on the new species of *Megaselia*. We thank Arina Adom for help in curating the Illumina barcodes and Sabrina Tang for help with photography.

REFERENCES

1. Wilson EO. Biodiversity research requires more boots on the ground. Nat Ecol Evol. 2017;1 11:1590-1. doi:10.1038/s41559-017-0360-y.
2. Krell F-T. Parataxonomy vs. taxonomy in biodiversity studies – pitfalls and applicability of ‘morphospecies’ sorting. Biodiversity and Conservation. 2004;13 4:795-812. doi:10.1023/b:Bioc.0000011727.53780.63.
3. Hebert PD, Cywinska A, Ball SL and deWaard JR. Biological identifications through DNA barcodes. Proceedings of Royal Society B. 2003;270 1512:313-21.
4. Wang WY, Srivathsan A, Foo M, Yamane SK and Meier R. Sorting specimen-rich invertebrate samples with cost-effective NGS barcodes: Validating a reverse workflow for specimen processing. Molecular Ecology Resources. 2018;18 3:490-501.
5. Meier R, Wong W, Srivathsan A and Foo M. $1 DNA barcodes for reconstructing complex phenomes and finding rare species in specimen-rich samples. Cladistics. 2016;32 1:100-10.
6. Yeo D, Puniamoorthy J, Ngiam RWJ and Meier R. Towards holomorphology in entomology: rapid and cost-effective adult–larva matching using NGS barcodes. Systematic Entomology. 2018;43 4:678-91.

7. Hebert PDN, Braukmann TWA, Prosser SWJ, Ratnasingham S, deWaard JR, Ivanova NV, et al. A Sequel to Sanger: amplicon sequencing that scales. BMC Genomics. 2018;19:219.

8. Shokralla S, Porter TM, Gibson JF, Dobosz R, Janzen DH, Hallwachs W, et al. Massively parallel multiplex DNA sequencing for specimen identification using an Illumina MiSeq platform. Scientific Reports. 2015;5:9687.

9. Creedy TJ, Norman H, Tang CQ, Chim KQ, Andujar C, Arribas P, et al. A validated workflow for rapid taxonomic assignment and monitoring of a national fauna of bees (Apiformes) using high throughput barcoding. BioRxiv. 2019; doi:10.1101/575308.

10. Krehenwinkel H, Kennedy SR, Rueda A, Lam A and Gillespie RG. Scaling up DNA barcoding – Primer sets for simple and cost efficient arthropod systematics by multiplex PCR and Illumina amplicon sequencing. 9. 2018;11 2181-2193.

11. Krehenwinkel H, Pomerantz A, Henderson JB, Kennedy SR, Lim JY, Swamy V, et al. Nanopore sequencing of long ribosomal DNA amplicons enables portable and simple biodiversity assessments with high phylogenetic resolution across broad taxonomic scale. GigaScience. 2019; doi:10.1093/gigascience/giz006.

12. Krehenwinkel H, Wolf M, Lim JY, Rominger AJ, Simison WB and Gillespie RG. Estimating and mitigating amplification bias in qualitative and quantitative arthropod metabarcoding. Scientific Reports. 2017;7 17668.

13. Srivathsan A, Baloglu B, Wang W, Tan WX, Bertrand D, Ng AHQ, et al. A MinION™-based pipeline for fast and cost-effective DNA barcoding. Molecular Ecology Resources. 2018;18 5:1035-49.

14. Brown BV: Phorid Catalog. http://phorid.net/pcat/. Accessed 27 April 2019.

15. Stork NE. How Many Species of Insects and Other Terrestrial Arthropods Are There on Earth? Annu Rev Entomol. 2018;63:31-45. doi:10.1146/annurev-ento-020117-043348.

16. Zhang ZQ. Animal biodiversity: An introduction to higher-level classification and taxnomic richness. Magnolia Press; 2011.

17. Forbes AA, Bagley RK, Beer MA, Hippee AC and Widmayer HA. Quantifying the unquantifiable: why Hymenoptera – not Coleoptera – is the most speciose animal order. BMC Ecology. 2018;18 21 doi:10.1101/274431.

18. Hebert PD, Ratnasingham S, Zakharov EV, Telfer AC, Levesque-Beaudin V, Milton MA, et al. Counting animal species with DNA barcodes: Canadian insects. Philos Trans R Soc Lond B Biol Sci. 2016;371 1702 doi:10.1098/rstb.2015.0333.

19. Borkent ART, Brown BV, Adler PH, Amorim DDS, Barber K, Bickel D, et al. Remarkable fly (Diptera) diversity in a patch of Costa Rican cloud forest: Why inventory is a vital science. Zootaxa. 2018;4402 1 doi:10.11646/zootaxa.4402.1.3.

20. Brown BV, Borkent A, Adler PH, De Souza Amorim D, Barber K, Bickel D, et al. Comprehensive inventory of true flies (Diptera) at a tropical site. Communications Biology. 2018;1 21:8. doi:10.1038/s42003-018-0022-x.

21. Marshall SA. Flies: The Natural History and Diversity of Diptera. Buffalo, New York: Firefly Books; 2012.

22. Brown BV and Hartop EA. 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 Ecosystems. 2016; doi:10.1007/s11252-016-0612-7.

23. Vaser R, Sovic I, Nagarajan N and Sikic M. Fast and accurate de novo genome assembly from long uncorrected reads. Genome Res. 2017;27 5:737-46. doi:10.1101/gr.214270.116.

24. Kwong S, Srivathsan A, Vaidya G and Meier R. Is the COI barcoding gene involved in speciation through intergenomic conflict?. Molecular Phylogenetics and Evolution. 2012;62 3:1009.

25. Hartop EA and Brown BV. 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. Biodiversity Data Journal. 2014;2:e4093. doi:10.3897/BDJ.2.e4093.

26. Hartop EA, Brown BV and Disney RHL. Opportunity in our ignorance: urban biodiversity study reveals 30 new species and one new Nearctic record for *Megaselia* (Diptera: Phoridae) in Los Angeles (California, USA). Zootaxa. 2015;3941:451-84.

27. Hartop EA, Brown BV and Disney RHL. Flies from L.A., The Sequel: Twelve further new species of *Megaselia* (Diptera: Phoridae) from the BioSCAN Project in Los Angeles (California, USA). Biodiversity Data Journal. 2016, p. e7756.

28. Riedel A, Sagata K, Surbakti S, Rene T and Michael B. One hundred and one new species of *Trigonopterus* weevils from New Guinea. Zookeys. 2013; 280:1-150. doi:10.3897/zookeys.280.3906.

29. ONT: https://store.nanoporetech.com/kits-250/1d-sequencing-kit.html. Accessed 28 April 2019.

30. Erwin TL. Tropical Forests: Their Richness in Coleoptera and Other Arthropod Species. The Coleopterists Bulletin. 1982;36 1:74-5.

31. Longino JT, Coddington J and Colwell RK. The ant fauna of a tropical rain forest: estimating species richness three different ways. Ecology. 2002;83:689-702.

32. Townes H. A light-weight Malaise trap. Entomological News. 1972;83:239-47.

33. Howard PC. Nature conservation in Uganda's tropical forest reserves. The IUCN Tropical Forest Programme. 1991.

34. Chapman CA and Chapman LJ. Forest regeneration in logged and unlogged forests of Kibale National Park, Uganda. Biotropica. 1997;29:396-412.

35. Kurina O. Description of four new species of *Zygomyia* Winnertz from Ethiopia and Uganda (Diptera: Mycetophilidae). African Invertebrates. 2012;53 1:205-20.

36. Folmer O, Black M, Hoeh W, Lutz R and Vrijenhoek R. DNA primers for amplification of mitochondrial cytochrome c oxidase subunit I from diverse metazoan invertebrates. Molecular Marine Biology and Technology. 1994;3 5:294-9.

37. Comai L and Howell T. Barcode Generator. *http://comailabgenomecenterucdavisedu/indexphp/Barcode_generator*. 2012.

38. Katoh K and Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Molecular Biology and Evolution. 2013;30 4:772-80.

39. Sovic I, Sikic M, Wilm A, Fenlon SN, Chen S and Nagarajan N. Fast and sensitive mapping of nanopore sequencing reads with GraphMap. Nature Communications. 2016;7:11307.

40. Vaser R, Sovic I, Nagarajan N and Sikic M. Fast and accurate de novo genome assembly from long uncorrected reads. Genome Research. 2017;27 5:737-46.

41. Leray M, Yang JY, Meyer CP, Mills SC, Agudelo N, Ranwez V, et al. A new versatile primer set targeting a short fragment of the mitochondrial COI region for metabarcoding metazoan diversity: application for characterizing coral reef fish gut contents. Frontiers in Zoology. 2013;10:34.

42. Geller JM, C., Parker M and Hawk H. Redesign of PCR primers for mitochondrial cytochrome c oxidase subunit I for marine invertebrates and application in all-taxa biotic surveys. Molecular Ecology Resources. 2013;13 5:851-61.

43. Yeo D, Srivathsan A and Meier R. Mini-barcodes are more suitable for large-scale species discovery in Metazoa than full-length barcodes. bioRxiv. 2019. doi:10.1101/594952.

44. Wang WY, Srivathsan A, Foo M, Yamane SK and Meier R. Sorting specimen-rich invertebrate samples with cost-effective NGS barcodes: Validating a reverse workflow

for specimen processing. Molecular ecology resources. 2018;18 3:490-501. doi:10.1111/1755-0998.12751.

45. Zhang J, Kobert K, Flouri T and Stamatakis A. PEAR: a fast and accurate Illumina Paired-End reAd mergeR. Bioinformatics. 2014;30 5:614-20.

46. Srivathsan A, Sha JC, Vogler AP and Meier R. Comparing the effectiveness of metagenomics and metabarcoding for diet analysis of a leaf-feeding monkey (*Pygathrix nemaeus*). Mol Ecol Resour. 2015;15 2:250-61. doi:10.1111/1755-0998.12302.

47. Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, et al. Versatile and open software for comparing large genomes. Genome biology. 2004;5 2:R12. doi:10.1186/gb-2004-5-2-r12.

48. Meier R, Shiyang K, Vaidya G and Ng PKL. DNA Barcoding and Taxonomy in Diptera: A Tale of High Intraspecific Variability and Low Identification Success. Systematic Biology. 2006;55 5:715–28.

49. Ahrens D, Fujisawa T, Krammer HJ, Eberle J, Fabrizi S and Vogler AP. Rarity and Incomplete Sampling in DNA-based Species Delimitation. Systematic Biology. 2016;65 3:478-94.

50. Leigh JW and Bryant D. PopART: Full-feature software for haplotype network construction. Methods Ecol Evol. 2015;6 9:1110-6.

51. Disney RHL. Scuttle flies (Diptera: Phoridae) Part II: the genus *Megaselia*. Fauna of Arabia. 2009;24:249-357.

52. Colwell RK. EstimateS: Statistical estimation of species richness and shared species from samples. Version 9 and earlier. User's Guide and application. 2013.

Figure Legends

Figure 1: Flowchart for generating MinION barcodes.

Figure 2: Effect of re-pooling on coverage of barcodes for both sets of specimens.

Figure 3: Ambiguities in MAFFT+AA (Purple), RACON+AA (Yellow) and Consolidated barcodes (Green) with varying namino parameters (1,2 and 3). One outlier value for Racon+3AA barcode corresponding to 13% ambiguities was excluded from the plot.

Figure 4: The Malaise trap that revealed the estimated >1000 mOTUs as shown by the species richness estimation curve. Green: Chao1 Mean, Pink: S (Mean), Orange: Singleton Mean, Purple: Doubleton mean.

Figure 5: Lateral habitus (a) and diagnostic features of *Megaselia sepsioides* spec. nov. (a, inset) terminalia, (b) posterior view of foreleg, (c) anterior view of midleg (d,e) anterior and postero-dorsal views of hindleg, (e) dorsal view of thorax and abdomen.

Figure 6: Haplotype variation of Megaselia sepsioides spec. nov. (a) UGC0005996, (b) UGC0012244, (c) UGC0012899. UGC numbers refer to specimen IDs.

Figure 7: Haplotype network for Megaselia sepsioides spec. nov. UGC numbers refer to specimen IDs.

Table 1: Number of reads and barcodes generated via MinION sequencing.

	Set 1: Two flowcells	**Set 2: One flowcell**	**Combined (set 1 & 2)***
# Specimens	4275	4519	8699
Resequencing (re-pooled)	2172	2211	
# reads/# reads >600 bp	7,035,075/3,703,712	7,179,121/2,652,657	NA
Initial sequencing (all)	3,069,048/1,942,212	4,853,363/2,250,591	
Resequencing (re-pooled)	3,966,027/1,761,500	2,325,758/402,066	
# demultiplexed reads	898,979 (24.3%)	647,152 (24.4%)	
Initial sequencing (all)	562,434 (29%)	561,383 (24.9%)	
Resequencing (re-pooled	336,545 (19%)	85,769 (21.3%)	
Combined results of original and resequencing runs			
# specimens with >=5X coverage	4227 (98.9%)	4287 (94.9%)	8428 (96.9%)
# MAFFT barcodes <1% N's	3797 (88.8%)	3476 (76.9%)	7220 (83%)
# MAFFT + AA barcodes	3774 (88.3%)	3464 (75.7%)	7178 (82.5%)
# RACON barcodes	3797 (88.8%)	3476 (76.9%)	7220 (83%)
# RACON +AA barcodes	3790 (88.7%)	3469 (76.7%)	7194 (83%)
# Consolidated barcodes	3740 (87.4%)	3394 (75%)	7115 (81.8%)
# Consolidated barcodes (non-phorids removed)	3700 (86.7%)	3369 (74.6%)	7062 (81.2%)
# mOTUS (2/3/4%)			**728/685/636**

* one plate was accidentally sequenced in both runs, wherever duplicates were present, second run was selected.

Table 2. Accuracy of MinION as assessed by Illumina barcodes. The overall optimal strategy is “Consolidated (namino=2)”. Optimal congruence values are highlighted in green.

Dataset	# compared with Illumina	Accuracy	# barcodes with errors/# >3% errors	mOTU richness deviation between MinION and Illumina barcodes		
				2%	3%	4%
MAFFT	6330	99.6091	4473/31	-3 (-0.43%)	**-1 (-0.15)**	-8 (-1.3)
RACON	6330	99.5075	4526/37	15 (2.08%)	5 (0.75)	**-3 (-0.48)**
MAFFT + AA (namino=1)	6146	99.9689	269/30	-6 (-0.85%)	-4 (-0.61)	-15 (-2.7)
MAFFT + AA (namino=2)	6146	99.9795	218/28	-8 (-1.15%)	-6 (-0.91)	-13 (-2.2)
MAFFT + AA (namino=3)	6286	99.967	350/28	-7 (-0.1%)	-4 (-0.61)	-14 (-2.2)
RACON + AA (namino=1)	6319	99.9617	419/27	5 (0.72%)	**1 (0.15)**	**-3 (-0.48)**
RACON + AA (namino=2)	6319	99.9711	326/27	2 (0.29%)	-2 (-0.31)	-5 (-0.81)
RACON + AA (namino=3)	6319	99.9664	383/29	4 (0.57%)	-2 (-0.3)	-8 (-1.28)
Consolidated (namino=1)	6230	99.9855	191/25	-1 (-0.14%)	-2 (-0.30)	**-3 (-0.48)**
Consolidated (namino=2)	6255	**99.9864**	**185/25**	**0 (0%)**	-3 (-0.45)	-4 (-0.64)
Consolidated (namino=3)	6250	99.9762	329/26	1 (0.14%)	**-1 (-0.15)**	**-4 (-0.48)**

Fig. 1

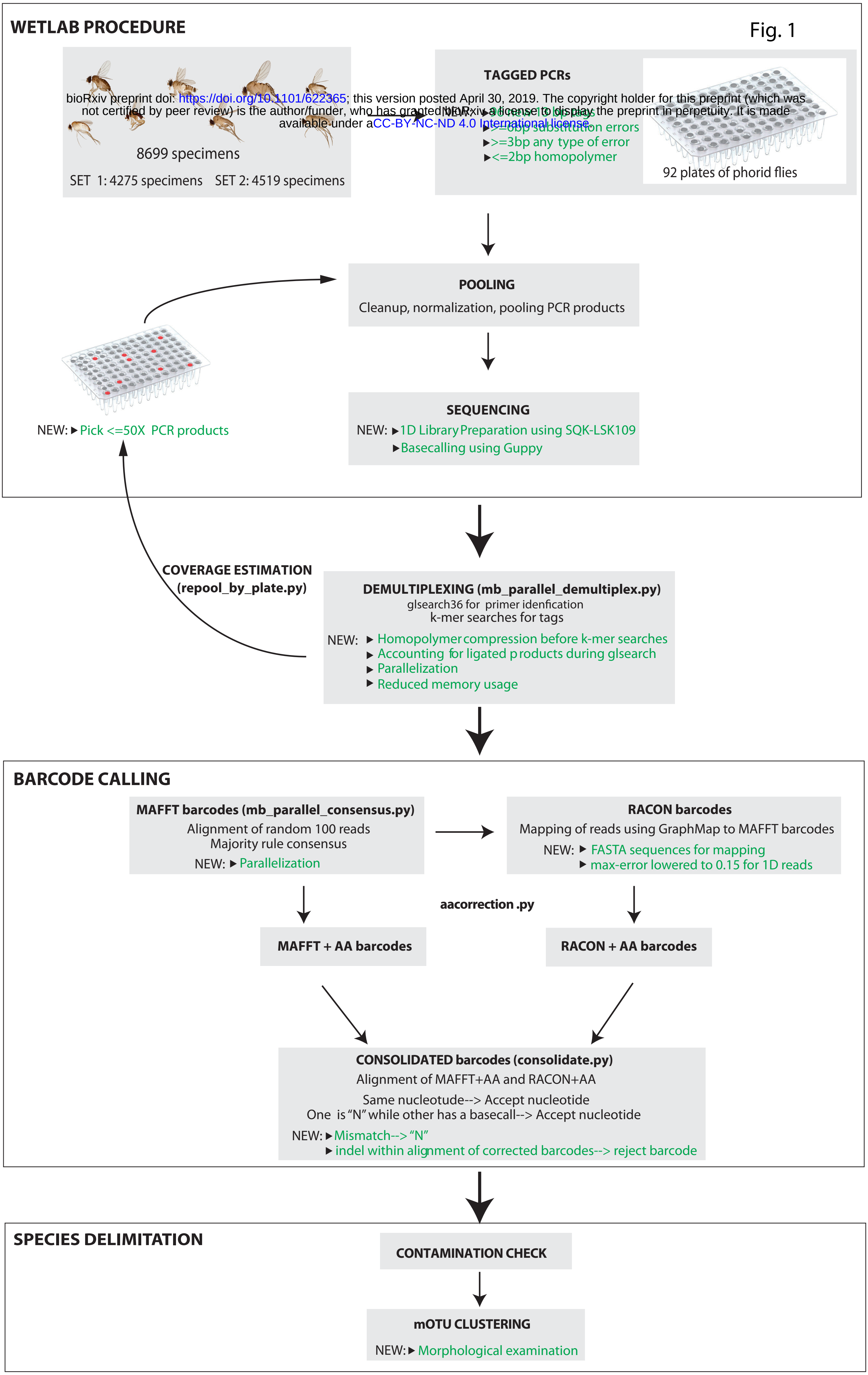
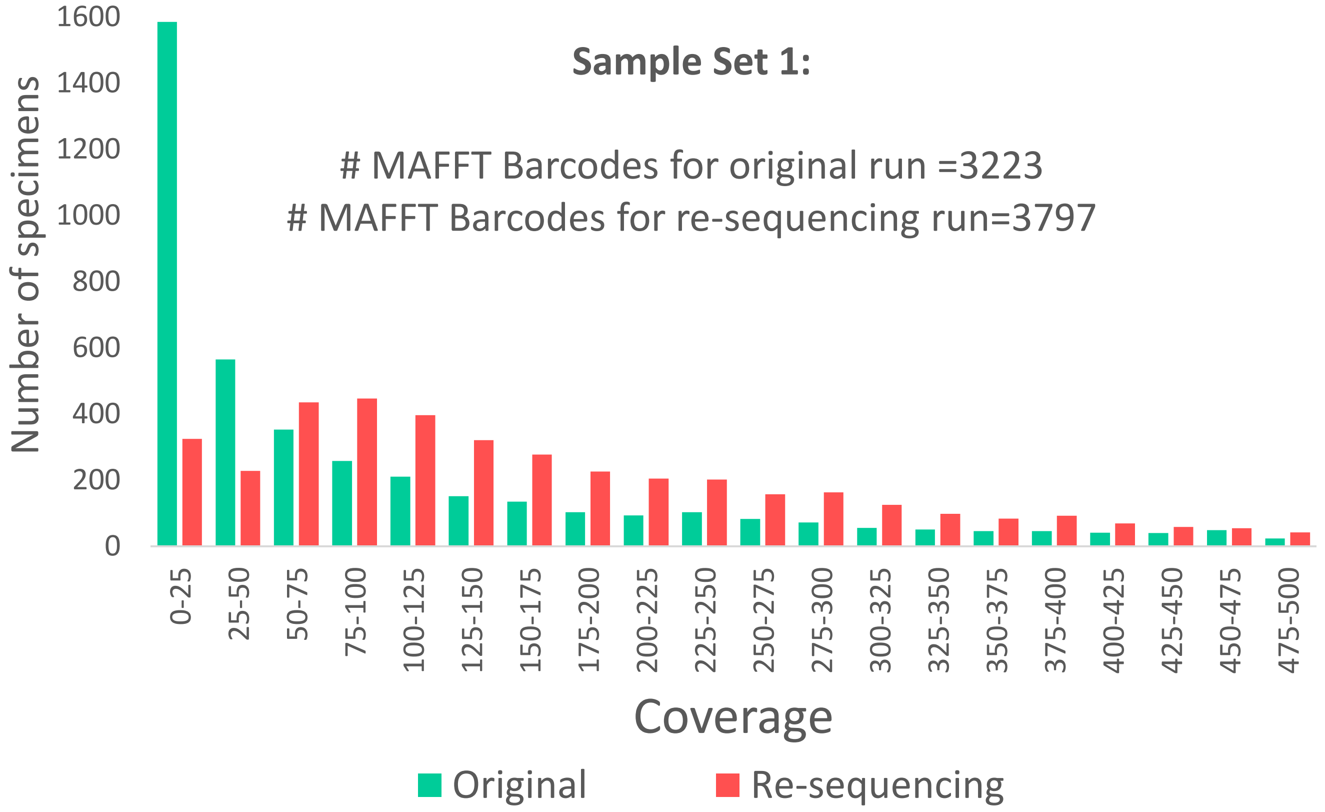

WETLAB PROCEDURE
8699 specimens
SET 1: 4275 specimens SET 2: 4519 specimens
TAGGED PCRs
NEW: ▶ 96 new 13 bp tags
▶ >=6bp substitution errors
▶ >=3bp any type of error
▶ <=2bp homopolymer
92 plates of phorid flies
POOLING
Cleanup, normalization, pooling PCR products
NEW: ▶ Pick <=50X PCR products
SEQUENCING
NEW: ▶ 1D Library Preparation using SQK-LSK109
▶ Basecalling using Guppy
COVERAGE ESTIMATION
(repool_by_plate.py)
DEMULTIPLEXING (mb_parallel_demultiplex.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 (mb_parallel_consensus.py)
Alignment of random 100 reads
Majority rule consensus
NEW: ▶ Parallelization
RACON barcodes
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 (consolidate.py)
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

Fig. 2


Sample Set 1:
# MAFFT Barcodes for original run =3223
# MAFFT Barcodes for re-sequencing run=3797
Number of specimens
0
200
400
600
800
1000
1200
1400
1600
0-25
25-50
50-75
75-100
100-125
125-150
150-175
175-200
200-225
225-250
250-275
275-300
300-325
325-350
350-375
375-400
400-425
425-450
450-475
475-500
Coverage
Original
Re-sequencing

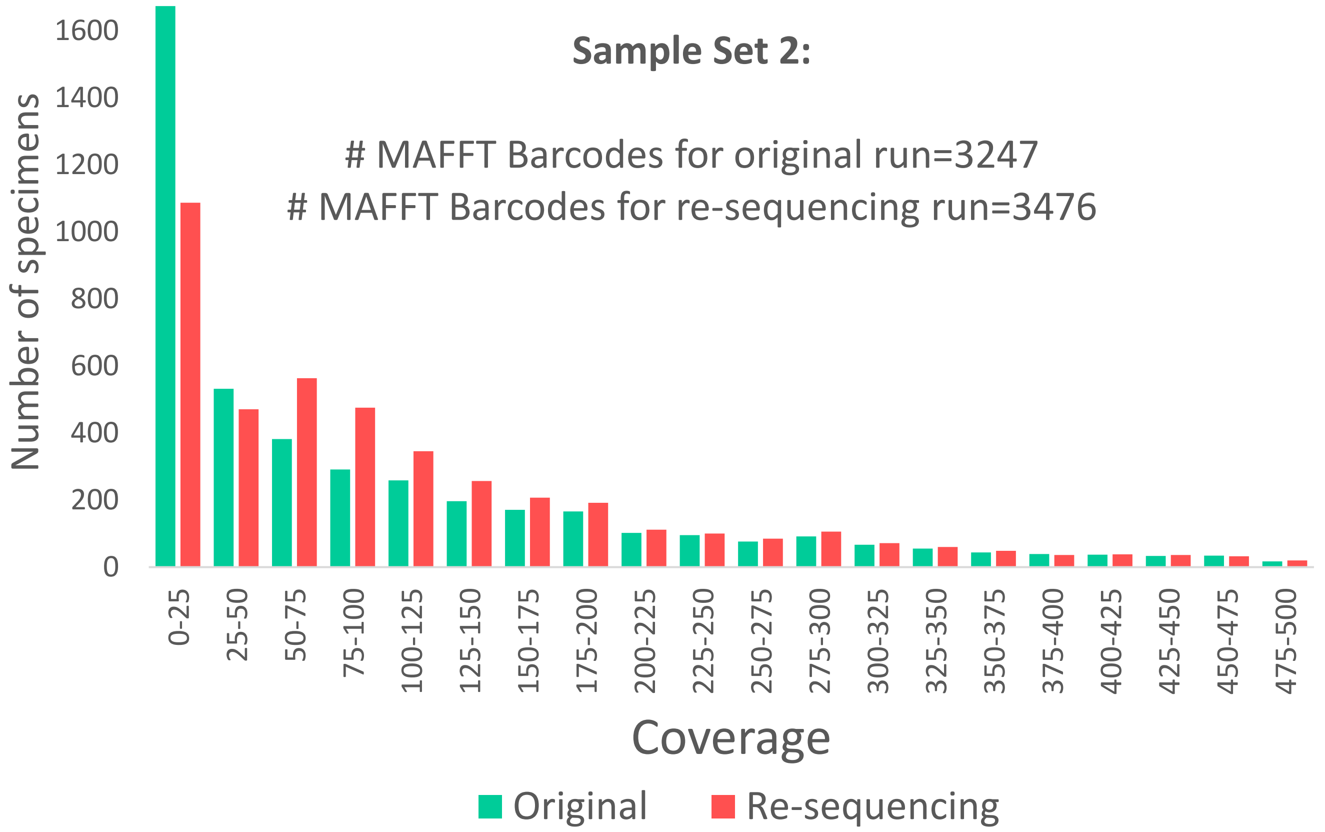

Sample Set 2:
# MAFFT Barcodes for original run=3247
# MAFFT Barcodes for re-sequencing run=3476
Number of specimens
0
200
400
600
800
1000
1200
1400
1600
0-25
25-50
50-75
75-100
100-125
125-150
150-175
175-200
200-225
225-250
250-275
275-300
300-325
325-350
350-375
375-400
400-425
425-450
450-475
475-500
Coverage
Original
Re-sequencing

Fig. 3

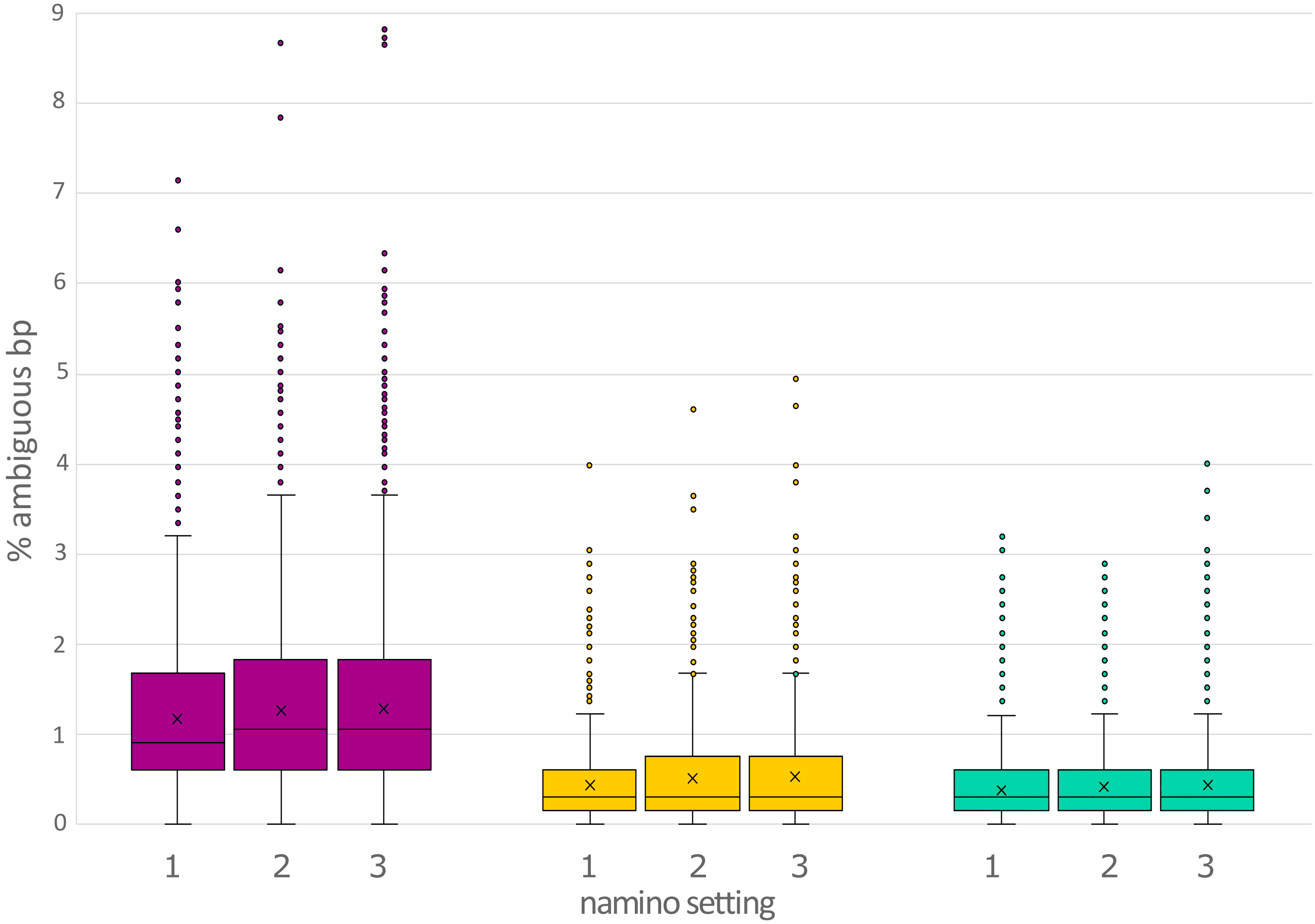


Fig. 4

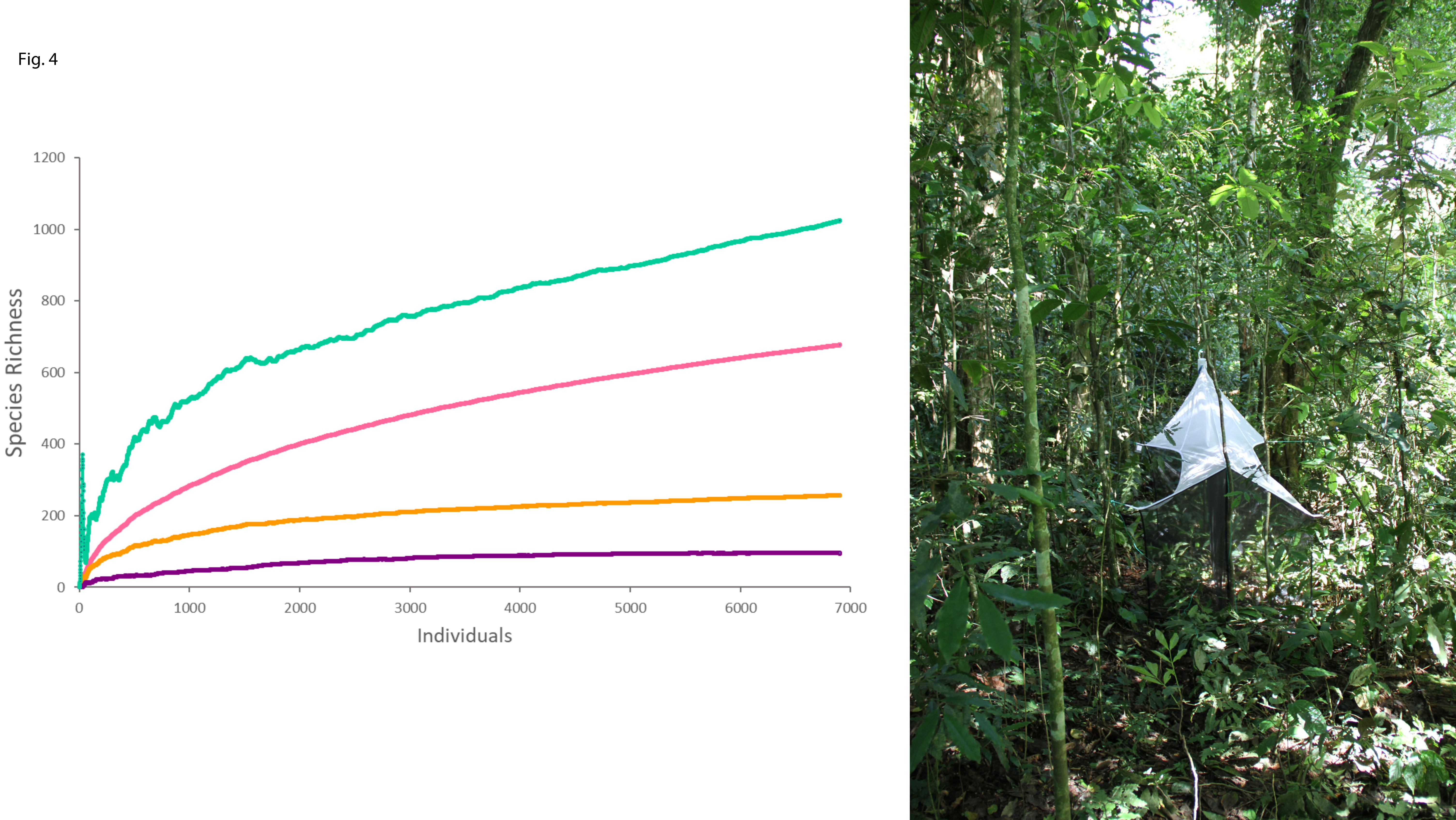

1200
1000
800
600
400
200
0
Species Richness
0
1000
2000
3000
4000
5000
6000
7000
Individuals

Fig. 5

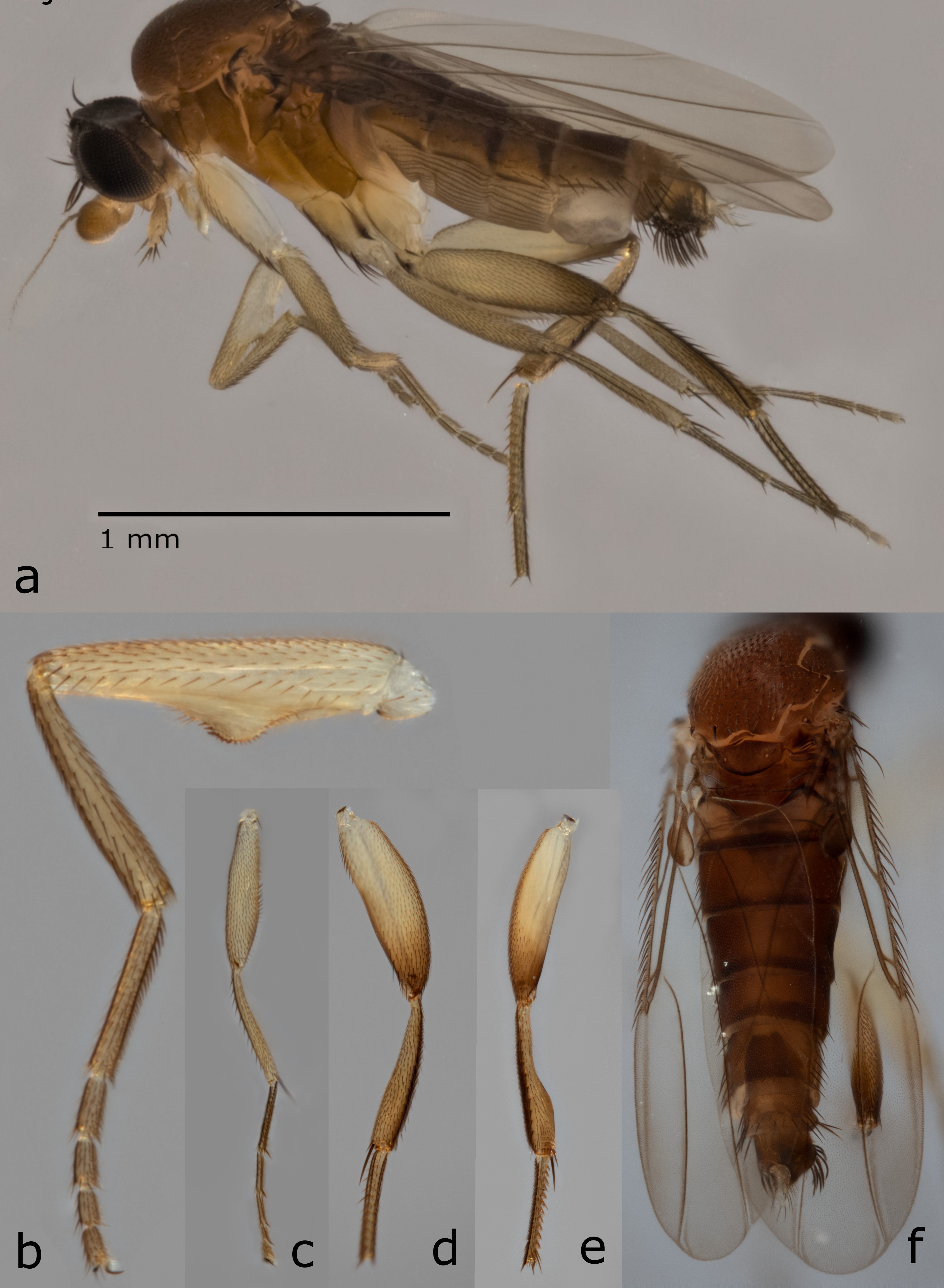


Fig. 6

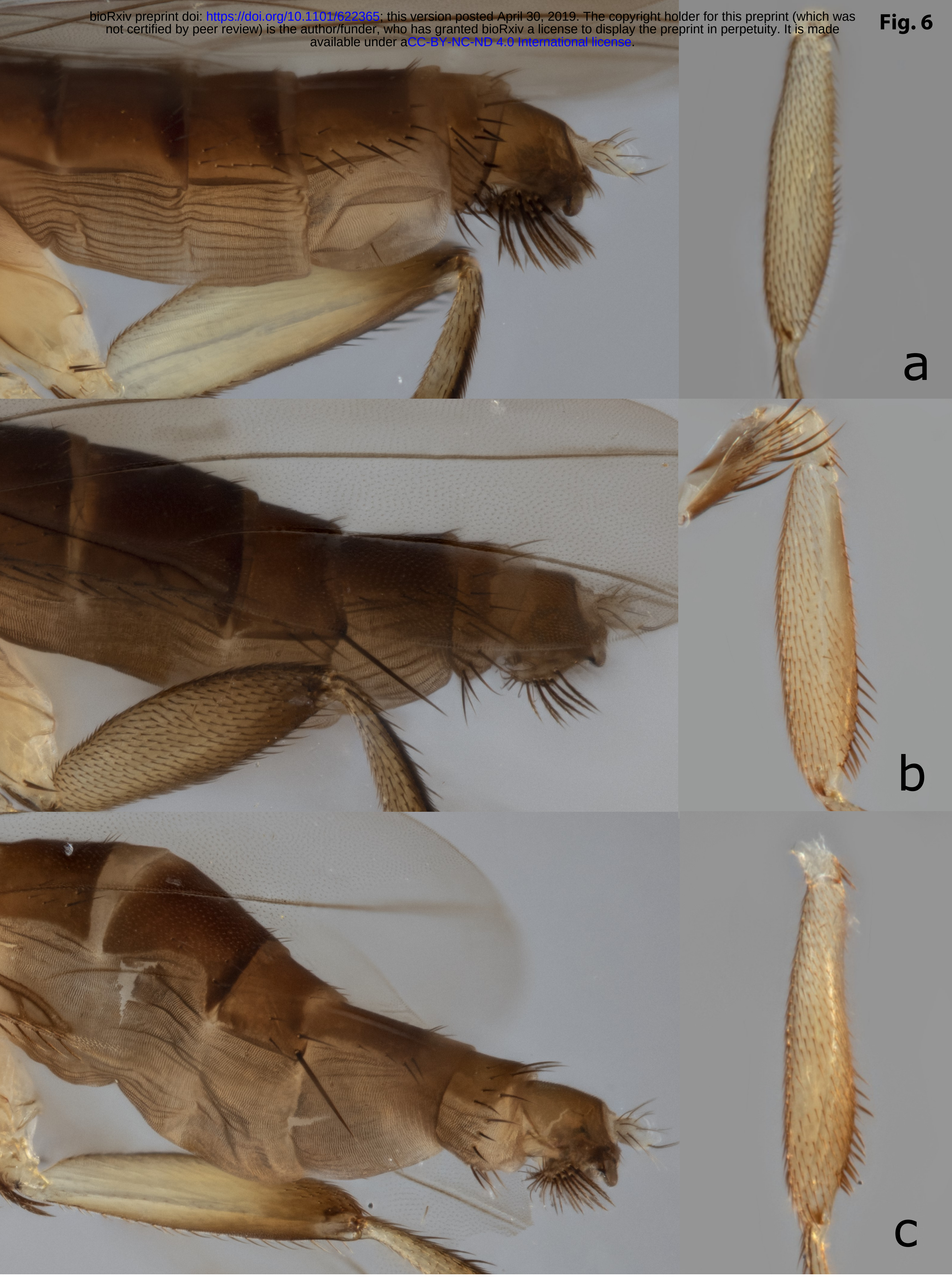
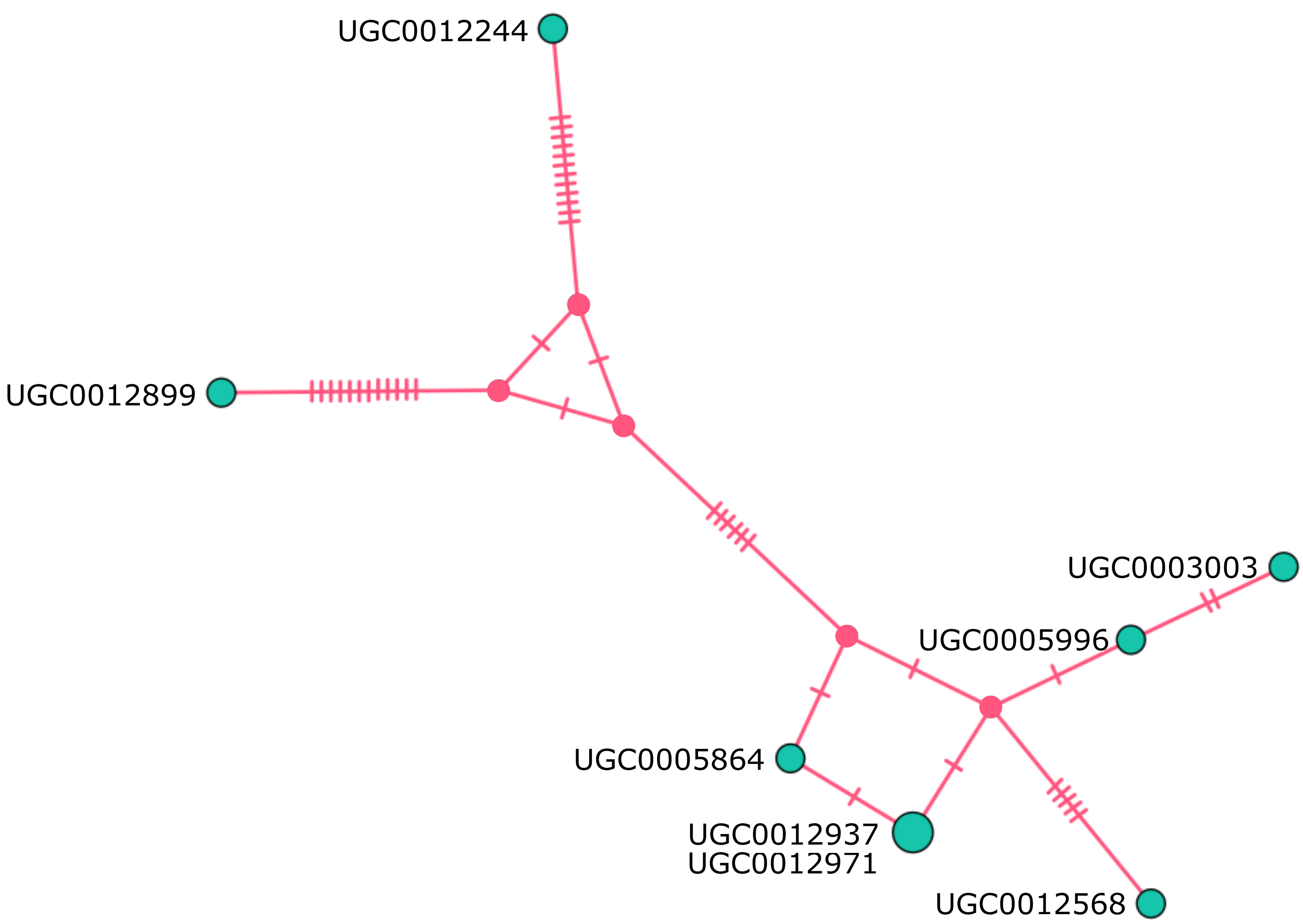


Fig. 7


UGC0012244
UGC0012899
UGC0003003
UGC0005996
UGC0005864
UGC0012937
UGC0012971
UGC0012568