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

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

ABSTRACT

Background:

Most animal species await discovery. Most of these undiscovered species share two features: they live in tropical regions and they belong to animal clades that combine high specimen abundance with immense species richness. Using morphology for species discovery in such taxa is slow because large numbers of specimens must be sorted to species based on detailed microscopic work. Fortunately, species discovery can be greatly accelerated by using molecular methods for sorting specimens into putative species before validating the species with morphology by examining a subset of specimens. This requires cost-effective and low-tech DNA barcoding techniques given that well equipped molecular laboratories are not readily available in many biodiverse countries.

Results:

We here document how MinION can be used to reveal the extent of the undiscovered biodiversity in specimen-rich taxa. We sequenced 7,062 specimens belonging to Phoridae, a hyper-diverse family of flies (Diptera). All specimens were collected in a single Malaise trap in Kibale National Park, Uganda over the period of eight weeks. We discover >650 species which exceeds the number of phorid species currently described for the entire Afrotropical region. We were able to obtain these barcodes at low cost using MinION sequencing by increasing the barcoding capacity per flowcell to 3500 (i.e., 700% increase over existing protocols). This is achieved by adopting 1D sequencing, re-sequencing weak amplicons on a used flowcell, and increasing the bioinformatics speed via parallelization and improved demultiplexing. Comparison with Illumina data reveals that the MinION barcodes yield very similar putative species (match ratio: 0.989) which is due to the high accuracy of the MinION barcodes (99.99% accuracy, 0.4% Ns). Morphological examination of 100 mOTUs confirms good congruence with molecular clusters (93% of mOTUs; >99% of specimens) and reveals that 90% of the putative species belong to the neglected megadiverse genus *Megaselia*. Lastly, we demonstrate for one

species how the molecular data guides the description of new species (*Megaselia sepsioides* sp. nov.).

Conclusions

We conclude that MinION is very suitable for reliable, rapid, and large-scale species discovery in hyperdiverse taxa, especially in biodiverse countries with limited access to capital-intensive sequencing facilities.

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 this requires the study of large numbers of specimens. Typically, the process starts with sampling specimens with bulk trapping methods (e.g. Malaise traps, fogging, pitfall traps, flight intercept traps) that yield thousands of specimens per site. These bulk samples then need to be sorted: first to higher-level taxonomic groups by parataxonomists and then to species-level by taxonomic experts [2]. Once morpho-species have been obtained, they are often tested with 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 clades that they tend to be 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 likely because barcoding with Sanger sequencing is so expensive that only few specimens can be barcoded for the purpose of validating morpho-species [4].

An alternative approach to species discovery is the ‘reverse workflow’ of Wang et al. (2018), wherein every specimen in a sample is DNA barcoded without the destruction of the specimen [4-6]. The specimens are then grouped into putative species (molecular Operational Taxonomic Units: mOTUs) based on DNA barcodes. The morphological check of these putative species is completed after sequencing and involves taxonomic experts working on pre-sorted material. Their work is guided by haplotype networks that indicate the genetic distance between specimens and allow the taxonomist to rapidly check the accuracy of the molecular pre-sorting by studying a subset of specimens. Described species are then identified and new species can

be described. Additionally, such molecular pre-sorting associates males, females, and immature specimens belonging to the same species [6]. Barcoding all specimens in a sample is unrealistically expensive with traditional Sanger sequencing but with the advent of High Throughput Sequencing platforms like Illumina and PacBio, large-scale barcoding has become more cost effective [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 are simultaneously barcoded and sequencing usually must be outsourced (amplicons have to be shipped to sequencing facilities). In industrialized countries, this usually takes weeks but the wait and cost is considerably more onerous in many biodiverse countries that lack high-throughput sequencing facilities because the DNA has to be sent on ice to an overseas facility. It would thus be desirable to have alternative sequencing techniques that are fast, scalable, cost-effective, and require low capital expenditure. Overall, such a solution would also be particularly useful if barcoding has to be accomplished under field conditions and/or by citizen scientists [11-15].

Oxford Nanopore's MinION has the potential to help with achieving these goals. It is a low-cost and portable real-time sequencing device, which, however, still generates error-prone data (ca. 10-15% [16]). Therefore, its use and reliability for large-scale specimen barcoding remains poorly explored. A first step toward the efficient use of MinION for barcoding was the recent demonstration that 500 reliable DNA barcodes can be obtained on one flowcell of MinION using $1D^2$ sequencing ([17]: taxa sequenced: dolichopodid flies, ants, chironomid midges, etc.). This was the highest number of barcodes that was ever successfully multiplexed in a single MinION flowcell (compare with [11]). This scale was considered useful for many species' identification projects, but it was unlikely to be effective for large-scale species discovery where thousands of specimens must be barcoded. Furthermore, the experiment used $1D^2$ sequencing, which

requires complicated and time-consuming library preparation techniques and server computers for the computationally intensive base-calling. Here, we test whether the more straightforward, but less accurate 1D sequencing can be used for large-scale species discovery.

New techniques for large-scale species discovery are urgently needed in order to understand the planet's species-level biodiversity, which is concentrated in hyperdiverse clades of invertebrates often living in the tropics. A good example is Phoridae (Diptera). Currently, only 466 species of phorid flies have been recorded for the Afrotropical Region [18] although Henry Disney, a world expert on the family, has recorded 75 species of phorids in his suburban garden in Cambridge alone [19]. Similarly, the BioSCAN project in Los Angeles found 82 species in the city's backyards [19]. These numbers make it very likely that the Afrotropical fauna is very large and currently vastly undersampled. This is likely to be the case for many taxa in the Diptera. This order is one of the four hyper-diverse insect orders with the others being 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 [20]) with only ca. 1,000,000 currently being described [21]. Historically, Coleoptera have been considered the most species-rich order of insects, but this may have been due to the *inordinate fondness* of taxonomists for beetles given that recent studies suggest that Diptera and Hymenoptera may have larger numbers of species. For example, Forbes et al. [22] hypothesize that Hymenoptera contained more species than either Diptera or Coleoptera based on parasite host ratios for Microhymenoptera, but Diptera have recently been shown to also be surprisingly rich in a number of large-scale biodiversity studies. In a large barcoding study of Canadian insects, Hebert et al. [23] 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 [23]; 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. For example, 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 [24]. Only 41,001 specimens (a small fraction of the hundreds of thousands collected) could be studied by taxonomic experts [25]. The specimens that were examined revealed 4,332 species of Diptera, of which 800 belonged to Cecidomyiidae and 404 to Phoridae [25], the fly family of focus here.

Phoridae, or scuttle flies, is a worldwide family of true flies with approximately 4300 described species [18]. 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” [26]. 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. 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 rare species cannot be found without the microscopic study of thousands of specimens for which prodigious notes have to be taken. Many detailed drawings are prepared (for *Megaselia* drawings of male genitalia are essential) – often based on dissections and slide mounts. The process can take hours, days, or longer for a single species discovery. This traditional workflow thus 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 by relegating the sorting to species level of a sample to sequencing. 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 describe how we processed ~8700 specimens, obtained ~7000 accurate barcodes, and found >650 putative species (of which almost 90% were *Megaselia*). All this could be accomplished using a workflow that took less than a month.

RESULTS

1. 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 8,699 phorid flies were processed (Set 1: 4275; Set 2: 4519; 95 specimens were duplicated in both 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) for weak amplicons that cannot be reliably detected with commercial dyes. 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. [17]. 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).

2. Assessment of demultiplexing accuracy

We indirectly assessed the accuracy of the demultiplexing pipeline by testing whether reads were incorrectly demultiplexed to unused tag combinations. This happened for a very small proportion of reads (0.23%: 2,054 of 900,698 reads in set 1; 0.44%: 2,837 of 649,587 reads in set 2). Fortunately, such low rates of incorrect de-multiplexed reads are unlikely to cause problems given that the average coverage per amplicon was 210X (set 1) and 143X (set 2). Note that 37% and 69% of these incorrectly demultiplexed reads were due to one tag: GTCCAACTTCAGT. The edit distances between all tag-pairs are high (>=5 bp) so that it is currently unclear whether this was caused by a primer synthesis issue, systematic sequencing bias, or a wet lab problem (Supplementary Figure 1). Out of caution, we have provided four additional tag sequences that can be used as replacements (Supplementary Table 4).

3. Barcode calling

Demultiplexing all data and calling preliminary barcodes revealed 3,797 and 3,476 MAFFT barcodes with >=5X coverage and <1% ambiguous bases. These barcodes were subject to correction using RACON [27] which yielded the same number of barcodes. After combining the barcodes for the two sets of samples, we overall obtained 7,220 MAFFT and RACON barcodes. These preliminary barcodes still contained indel and substitution errors that were addressed with an amino-acid correction pipeline that was first implemented in Srivathsan et al. [17]. It yielded 7,178 MAFFT+AA and 7,194 RACON+AA barcodes. Note that the drop in barcode numbers was 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 to 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 be phorid flies (<1%). These included barcodes with matches to *Wolbachia, Rickettsia,* nematodes, human, and occasionally insects from other families (e.g. *Drosophila, Hemipyrellia*). After removal of these, we retained 7,062 barcodes for estimating the species richness at the Ugandan site. 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).

4. Comparison of MinION sequence with Illumina sequence

Illumina barcodes were obtained for 6,373 of the same specimens using different primers that amplified a 313 bp subset of the full-length barcodes; i.e. the following comparison is thus for 47% of the MinION sequence length, but we are not aware of any reasons why the remaining 53% of the sequence would have different error rates. 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 %). These accuracy rates were obtained after excluding the <0.5% specimens that had >3% divergence with corresponding Illumina barcodes. Such barcode discrepancies are likely due to wet-lab errors or due to the amplification of residual contaminating signals (see details in methods). Such errors are not unexpected for large-scale barcoding projects. For examples, a recent study by Hebert et al. [7] using PacBio Sequel for DNA barcoding found that 1.5-1.6% of the specimens had high abundances of non-target sequences.

5. Comparison of MinION and Illumina barcodes at mOTU-level

Given that the barcodes were obtained for the purpose of species richness estimates, we compared the mOTU richness estimated for the different barcode sets against those obtained with Illumina barcodes. For this purpose, we trimmed the MinION barcode sets to the 313 bp fragment that was sequenced using Illumina. 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 (Table 2: 661 vs 655 for 3% threshold): 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). The consolidated barcodes and RACON barcodes performed best with match ratios of >0.98 (consolidated barcodes: 0.989, RACON: 0.989, RACON+AA: 0.982). However, upon closer inspection the multiple sequence alignment (MSA) for these barcodes contained indels while the MSA for consolidated barcodes was indel-free. The largest number of indels was found in the MSA of uncorrected RACON barcodes which indicated that the RACON barcodes retained a fair number of indel errors and may not be of sufficient quality for submission to sequence databases. We would thus recommend the usage of consolidated barcodes as the best estimate of the specimen barcode. This recommendation is based on maximizing per-base accuracy (see below), yielding high-quality alignments, and revealing very similar mOTU diversity and composition (high match ratio) when compared to Illumina.

We also compared the mOTUs determined by full-length consolidated MinION barcodes (658 bp) with the mOTU diversity obtained with Illumina for those specimens for which both types of data were available. The match ratio was again high (0.953). For the clusters that were incongruent, we analysed at which threshold they would become congruent. We found that all clusters were congruent within the 1.9-4% range; i.e., the remaining 345 bp are not showing a major deviation from the signal obtained from the 313 bp fragment (Supplementary Table 1). We next characterized if there was an increase in error in the 345 bp stretch of the MinION sequence that could not be directly compared to Illumina sequence: if this were the case, we would expect that spurious base calls would increase genetic distances for specimens. However, we found the opposite: in 19 of 22 cases, the threshold was lowered, i.e., the 345 additional nucleotides reduced the minimum distance in the cluster (Supplementary Table 1) and it is likely that the reduction in match ratio from 0.98 to 0.95 is due to having analyzed a longer segment of COI.

6. Species richness estimation

We proceeded to characterize the diversity of phorid flies collected from the Malaise traps based on the consolidated barcodes (namino=2). We obtained a mean of 683 mOTUs when the thresholds were varied from 2-4% (2%: 728, 3%: 685, 4%: 636). These thresholds are widely used in the literature, but they were here also empirically derived using GenBank data. GenBank has 12,072 phorid sequences with species-level identifications and they belong to 106 species. The intraspecific variability is overwhelmingly under 3% (>95% of pairwise distances) and the match ratios between mOTUs and species identifications from GenBank are maximized for clustering thresholds of 2 - 3% (Supplementary Figure 2, 3).

Species accumulation and Chao 1 curves for the Uganda phorid mOTUs at 3% were not found to have reached a plateau, but the shape of the curves suggest an estimated diversity of >1000 species of Phoridae in a single site collected by one Malaise trap (Fig. 4).

7. Search for COI paralogs

In order to rule out that the high species richness estimate was due to the presence of paralogs, we scrutinized the Illumina read sets for co-occurrence of multiple COI signals in the same mOTU at trace levels. Illumina datasets were used because they were amplified with more degenerate primers that increase chances of co-amplification of paralogs. All mOTUs with multiple phorid signals in their read set either have paralogs that amplified or were affected by low-level template contamination between wells. Overall, we only found 30 of 406 mOTUs with multiple specimens containing such signals; i.e., even if all were caused by the presence of paralogs, the species richness estimates would remain >650 (Supplementary Table 3).

8. Congruence with morphology

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 of 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 [28]. This means that taxonomists working with mOTUs should check for signs of lumping and splitting in closely related taxa. This requires careful taxonomic examination of the specimens which is here aided by physically grouping the clusters based on genetic distances. Closely related mOTUs are kept physically together by clustering at larger threshold (5%) and sub-clusters are kept compartmentalized within this large cluster. This allows the taxonomist to focus on studying representatives of each sub-cluster. In the case of phorids this can be done by slide mounting representative specimen from the sub-clusters. Our initial morphological check of 100 randomly selected clusters (>1500 specimens) took ca. 30 hours. This 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 their 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* [29-31] and uses a barcode and photographic description. Photographs are a key element in descriptions for large, diverse groups [32], 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:

actttatattttattttttggagcttgagctggaatagtaggtacttccttaagaatcataattcgtgctgaattaggacacccaggagcacttat

tggtgatgaccaaatttataatgtgattgttactgcacatgcttttattataatttttttatagtaatacctattataataggaggttttggtaattg

acttgtacctttaatattaggagccccagatatggcattccctcgaatgaataatataagtttttgaatattacctccttctttaactcttttatta

gccagaagtatagtagaaaatggagctggaactggttgaacagtttatcctcctttatcttctagaatcgctcatagtggagcttctgttgat
ttagcaattttctctcttcatttagctggaatttcatctattttaggagctgtaaattttattacaacaattattaatatacgatcatcaggtattaca
tttgaccgaatacctctatttgtttgatctgtaggtattacagctttattgctactcttatcacttcctgttttagctggtgctattacaatactattaa
cagaccgaaattttaatacttcatttttgacccagcaggaggaggagatccaattttataccaacatttattc

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

Remarkably high diversity of Phoridae in Kibale National Park

We here reveal that a modest amount of sampling (one Malaise trap placed in Kibale National Park, Uganda) and the barcoding of 7000 specimens can lead to the discovery of >650 putative species of phorid flies. This diversity constitutes 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 and are likely new to science [18]. 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 to exceed 1000 observed species. This prediction is also supported by a formal species-richness estimation based on the available data (Fig. 4). Such extreme diversity from a single site raises the question of whether these numbers are biologically plausible and/or whether they could be caused by unreliable data or species delimitation.

As mentioned earlier, a single garden in Cambridge (UK) can have 75 species of phorids and the backyards of Los Angeles are known to contain up to 82 species. We are here observing ca. 10-15 times this diversity in a tropical National Park site and we would argue that this is biologically plausible given that tropical sites tend to have much higher species diversity than

temperate sites. In addition, the comparison sites in the UK and USA are urban. The likelihood that there would be 1000 species at a single site is further supported by the results of the Zurqui survey in Costa Rica which revealed 404 species of Phoridae without completing the species discovery process [24]. We are also confident that our species richness estimate is not an artefact of poor data quality because the high species richness is supported by both Illumina and MinION data generated independently from the same samples using different primer pairs. It is furthermore stable to modifications of clustering thresholds. These thresholds were obtained by analysing the available phorid data in GenBank. It revealed that the usual 2-4% species-level thresholds that are generally used across most of Metazoa are also appropriate for phorid flies [33]. Lastly, we checked 100 randomly chosen mOTUs for congruence between molecular and morphological clusters. We find that 93% of the clusters and >99% of specimens are congruently placed (six of the seven cases of incongruence involved single specimens). This is in line with congruence levels that we observed previously for ants and odonates [4, 6].

The high species richness found in one study site inspired us to muse about the species diversity of phorids in the Afrotropical region, much as Terry Erwin did when he famously estimated a tropical arthropod fauna of 30 million species based on his explorations of beetle diversity in Panama [34]. Such extrapolation is useful because it raises new questions and inspires follow-up research. Speculation is inevitable given that it remains remarkably difficult to estimate the species richness of diverse taxa [35]. 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 putative species which likely represent a species community 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 which is likely to be home to several thousand species of phorids. If we assume

that on average each of the two-thousand 100 km^2 squares have 100 endemic phorid species, the “endemic” phorids alone would contribute 200,000 species of phorids to the Afrotropical fauna without even considering the contributions by the remaining species with a wider distribution. Based on our sample from Kibale National Park, 90% of species and specimens would belong to *Megaselia* as currently circumscribed. These estimates would only be much lower if the vast majority of phorid species had very wide distributions. We consider this 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 taxonomic workflow, almost all *Megaselia* specimens would remain in the unsorted Malaise trap residues for decades or centuries. Indeed, there are thousands of vials labelled as “Phoridae” shelved in major museums worldwide. This is because the traditional species discovery workflows are not able to keep pace with the species numbers and specimen abundance. This is what 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 are currently considered unworkable with traditional techniques. We would argue that MinION barcodes will be one of the important methods that will be used to tackle these clades. This predication is based on the fact that MinION barcodes can be obtained without having to invest heavily into sequencing facilities, MinION laboratories can be mobile, and even operate under field conditions [12-15]. MinION barcodes will also be important for the “democratization” of biodiversity research because the data are generated quickly enough that they can be integrated into high school and citizen scientist initiatives.

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, re-pooling of low coverage amplicons, and an improved bioinformatics workflows, we are here increasing the barcoding capacity of a MinION flowcell by 700% from 500 specimens (Srivathsan et al., 2018: $1D^2$ sequencing) to ~3500 specimens. This is achieved without a drop in accuracy because the error correction pipeline is effective at eliminating most of the errors in the 1D reads (ca. 10%). Indeed, even the initial estimates of barcodes ("MAFFT" & "RACON") have very high accuracy (>99.5%) when compared to Illumina data, while the accuracy of consolidated barcodes is even higher (>99.9%). These accuracy values are comparable to the accuracy of barcodes obtained by large-scale barcoding using PacBio Sequel [7: Additional File 1: Figure S5]. Note that the accuracy of the barcodes obtained in this study are even higher than what was obtained with $1D^2$ sequencing in Srivathsan et al. (2018: 99.2%) [17]. 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) [36].

We tested a range of different techniques for obtaining barcodes from 1D reads which raises the question of which bioinformatics pipeline should be used. 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 when $1D^2$ reads were used previously [17]. However, we find that all barcode sets yield reliable mOTU estimates; i.e., the higher error rates of the uncorrected barcodes rarely affected mOTU delimitation. 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).

Based on the procedures described here, MinION barcodes can be generated rapidly and at a low sequencing cost of <0.35 USD per barcode. Molecular cost of PCR is 0.16 USD per reaction [4] while QuickExtract™ reagent costs 0.06. These properties make MinION a valuable tool for species discovery whenever a few thousand specimens must be sorted to species (<5000). Larger-scale barcoding is probably still best tackled with Illumina short-read or PacBio's Sequel sequencing [4, 6, 7] because the barcoding cost is even lower. However, both require access to expensive sequencers, the sequencing must 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. Another advantage of our proposed MinION pipeline is that it only requires basic molecular lab equipment including thermocyclers, a magnetic rack, a Qubit, a custom-built computational device for base-calling ONT data ("MinIT"), and a laptop (cost <USD 10,000). Arguably, the biggest operational issue is 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 are then re-sequenced in the second run. This dual-run strategy overcomes 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 pool of weak amplicons will contain sufficient DNA. This ability to re-pool within days of obtaining the first sequencing results is a key advantage of MinION. The same strategy can be applied to Illumina and PacBio but it would take a long time to obtain the results of two consecutive runs.

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 consistency with morphology. Reverse-workflow 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. Arguably, expert sorting and identification of material based on morphology is only efficient for taxa where species-specific morphological characters are easily accessible and/or the number of specimens is small. However, many 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 molecular sorting is carried out by lab personnel trained in amplicon sequencing while accurate morpho-species sorting requires highly specialized taxonomic experts. Yet, even highly trained taxonomic experts are often not able to match the males and females belonging to the same species (often one sex is ignored in morphological sorting) while the matching of sexes (and immatures) is an automatic and desirable by-product of applying the reverse workflow [6].All these benefits can be reapt rapidly: we estimate that the species-level sorting of 8000 specimens can be completed in less than one month given that one lab member can sequence 600-1000 specimens per day. Such a study would require two primary MinION runs, two secondary MinION runs, with two days of pooling weak products. This would still leave more than 1 week for clustering the specimens.

One key element of the reverse workflow is that vials with specimens that have haplotype distances <5% are physically kept together. This helps when assessing the congruence between mOTUs and morphology. Indeed, graphical representations of haplotype relationships (e.g., haplotype networks) are the guide for the 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 the presence of one or two species. Variations in setation were observed (Fig. 6) but they were consistent with the intraspecific level obtained in other phorid species. 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 accessible because taxonomic experts can work on pre-sorted material instead of thousands of unsorted specimens.

METHODS

1. Sampling

Samples were collected from a single Townes-type Malaise trap [37], 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 [38]. Temperatures in Kibale range from 16°C to 23°C (annual mean daily minimum and maximum, respectively) [39]. The Malaise trap was checked every week when the collecting bottle with the material was replaced by a resident parataxonomist ([40]: Mr. Swaibu Katusabe). Subsequently, the material 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 the phorid fraction of the total sample. 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 whole flies. The fly first taken out from the vial and washed in Milli-Q® water prior to being placed in a well of a 96 well PCR plate. DNA extraction was done using 10 ul of QuickExtract™ (Lucigen) in a 96 well plate format and the whole fly was used to extract DNA. The reagent allows for rapid DNA extraction by incubation (no centrifugation or columns

are required). 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' [41]. This primer pair has had high PCR success rates for flies in our previous study [17] and hence was chosen for phorid flies. 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 BarcodeGenerator (http://comailab.genomecenter.ucdavis.edu/index.php/Barcode_generator) [42] However, tag distances of >6 bp are not sufficiently distinct because they do not take into account MinION's propensity for creating errors in homopolymer stretches and other indel errors. We thus excluded tags with homopolymeric stretches that were >2 bp long. We next used a custom script to identify tags that differed from each other by indel errors. Such tags were eliminatedrecursively 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 procedure yielded a tag set with the edit-distance distribution shown in Supplementary Figure 1 [minimum edit distance of 5 nucleotides (38.5%) which is much higher than Nanopore error rates]. The edit (Levenshtein) distances were calculated by *stringdist* module in Python. Lastly, we excluded

tags that ended with “GG” because LCO1490 starts with this motif. Note that longer tags would allow for higher demultiplexing rates, but our preliminary results on PCR success rates suggested that the use of long tags reduced amplification success (one plate: 7% drop).

The PCR conditions for all amplifications were as follows, reaction mix: 10 µl Mastermix (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 (5 min). 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 inadvertently duplicated across the two runs.

II. MinION sequencing

We developed an optimized strategy for nanopore sequencing during the study. For the initial experiment (set 1), we sequenced amplicons for 4275 phorid flies. For this, all plates were grouped by amplicon strength as judged by the intensity of products on agarose gels and pooled accordingly (5 strong pools +2 weak pools). The pools were cleaned using either 1X Ampure beads (Beckman Coulter) or 1.1X Sera-Mag beads (GE Healthcare Life Sciences) in PEG and quantified prior to library preparation. 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 products with low coverage (<=50X), prepared a new library and sequenced them on a new flowcell. The experiment was successful. However, in order reduce sequencing cost and improve initial success rates, we pursued a different strategy for the second set of specimens (4519 specimens). Firstly, each plate was pooled separately for this experiment. In the first sequencing run, we stopped the sequencing 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 then 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.

Pooling of weak products with <=50X coverage was done as follows: 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 microplates that illustrate the wells where the weak products are found (available in github repository for *miniBarcoder*).

Library preparation and sequencing: We used the SQK-LSK109 ligation sequencing kit (Oxford Nanopore Technologies) 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 (Beckmann Coulter) 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 attempted to demultiplex the data for set 1 using *minibar* [11], however it was found to demultiplex only 1039 specimen barcodes for the 4275 specimens (command used: python ../minibar/minibar.py -F -C -e 2 minibar_demfile_1 phorid_run1ab.fa_overlen599 > out). This success rate was so low that we discontinued the use of *minibar*. Instead, we analysed the data using an improved version of *miniBarcoder* [17]. This pipeline starts with a primer search with *glsearch36,* followed by identifying the "sequence tags" in the flanking nucleotide sequences, before the reads are demultiplexed based on tags. For the latter, errors of up to 2 bp are allowed. These "erroneous" tags are generated by "mutating" the original tags 96 tags to account for all possible insertions/deletions/substitutions. The sequence tags are matched with this set of input tags + tag mutants. This speeds up the demultiplexing algorithm as it does not have to align each tag. When we compared the performance of *miniBarcoder* with *minibar*, we found that using 4 cores, *miniBarcoder* could demultiplex data in <10 minutes while *minibar* required 74 minutes in one. Both demultiplexed similar numbers of reads: 898,979 reads using *miniBarcoder*, while 940568 HH reads of *minibar* (56648 reads in multiple samples). The demultiplexed reads are aligned using MAFFT v7 (--op 0) (here v7) [43]. In order to improve speed, we only use 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".

Other studies usually incorporate a step of clustering of data at low thresholds (e.g. 70% by Maestri et al. 2019 [15]) to account for read error in MinION. The subsequent analysis is carried out on the cluster that has the larger number of reads. We deviate from this approach because it

does not exclude contaminated products and requires higher coverage. In our barcoding pipeline, we assess congruence for each basepair by firstly eliminating the MAFFT gap-opening penalty (--op 0); this allows for all data to be used when calling the consensus [17]: this gap opening penalty essentially treats indel and substitutions similarly and staggers the alignment. Any base that is found in <50% of the position is called as an ambiguity. We prefer that such a 50% criterion is applied at the character- instead of the read level which averages the impacts. The character-based approach maximizes data use by allowing good barcodes to be called at much lower coverage levels than with other pipelines (5-20X coverage compared to >200x coverage in Maestri et al. 2019 [15]), while also identifying contaminated specimen barcodes.

In our pipeline, MAFFT barcodes are improved by mapping all reads back to the barcode using Graphmap (here v0.5.2) [44] and calling the consensus using RACON (here, v1.3.1) [27]. 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 the 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. [17]). 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. [17] was modified as follows:

a. *Tackling 1D reads for massively multiplexed data:* We developed ways for correcting for the increased number of errors in 1D reads by identifying objective ways for quality assessments based on the MinION data and publicly 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 very long reads that can be formed when two amplicons ligate during library preparation. 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 <=20,000 sequences. The latter modification is needed given the rapid increase in MinION output and batch processing is scalable with increased output.

4. Assessment of demultiplexing accuracy

Demultiplexing accuracy was assessed by using tag combinations in the demultiplexing files that were not used in this experiment: 96x96 tags combinations were designed and set 1 used only 45 tags with LCO1490. This allowed for an assessment if *miniBarcoder* bins erroneously contained any reads for the remaining 51X96 combinations. Set 2 used 48 tags thus allowing us to assess if the demultiplexing pipeline erroneously found the remaining 48X96. Lastly one plate used only 55 tags associated with HCO2198. Overall, we could assess demultiplexing accuracy associated with 80/96 tags.

5. Illumina Based NGS Barcoding for validation

In order to validate MinION barcodes and optimize 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 m1COIintF: 5'-GGWACWGGWTGAACWGTWTAYCCYCC-3' [45] and modified jgHCO2198: 50-TANACYTCNGGRTGNCCRAARAAYCA-3 [46]. We recently conducted an extensive analysis to understand if a 313 bp minibarcode is able to provide similar

identifications and species delimitations as 658 bp barcodes and found that minibarcodes performing equally well as full-length barcodes when examining >5000 species [47].

Tagged primers were used for the PCRs as specified in Wang et al. (2018) [4]. 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) [4]: paired end reads were merged using PEAR (v 0.9.6) [48], 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 with >50x dataset coverage and 10 X coverage for the sequence, and a barcode is accepted only if this sequence is 5X as common as the next most common sequence. Demultiplexed reads with >10x coverage were also per sample were retained for analyses of contaminations and paralogs.

6. 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* [49]. 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 [50].

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 to be caused by wetlab error whenever a different product as the Illumina and MinION barcodes were amplified using different primers (<0.5% of specimens) as discussed here: for consolidated barcodes 25/6255 specimens are involved in such >3% distances in the comparisons of consolidated barcodes with Illumina barcodes. If these distances were a result of consensus errors, sequences are unlikely to create consensus that are identical to other specimens. 20/25 specimens have identical sequences to some other specimen and two are within 0.6% of another. By matching MinION barcodes to Illumina's read set for each sample (>=10 counts), we also find that many of the signals being picked up by HCO-LCO are in the Illumina data at low frequencies (in 14/25 cases, including 1 of the 3 cases that remain unaccounted) that was generated using different primers (Supplementary Table 2). We do not include these for consensus quality, but instead consider them as misplaced sequences.

We were furthermore 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) [51]. mOTU richness was estimated and we furthermore calculated match ratio between two sets of clusters[52]. Match ratio is given by $\frac{2N_{match}}{N_1+N_2}$. For consolidated barcodes, we also calculated the match ratio between mOTU estimates by full length barcodes and Illumina barcodes (313 bp). If the mOTUs were inconsistent we also identified the distance threshold for which they became consistent.

7. Search for COI paralogs

We next searched for nuclear copies of mitochondrial sequences (NuMTs), or paralogs. Our first check for paralogs is the comparison between MinION and Illumina barcodes that were generated using different primer pairs. Our next test involved the use of Illumina datasets as these are amplified using highly degenerate primers, which make it more likely that they would co-amplify paralogous copies along with the barcode sequence. We started with the demultiplexed read set for each specimen that was used for the comparison with the consolidated barcode set of MinION as species richness estimates based on this set of specimens and MinION barcodes were very similar (match ratio >0.98, <0.5% specimens misplaced). For each demultiplexed set, we obtained the unique sequences that are >=10X coverage. All the secondary sequences (sequences that are not barcode sequences) were matched against the barcode sequences using BLASTN, e-value 1e-5, perc_identity 99%, hit length >=300. If secondary signals for a specimen belonged to a different specimen these could be low level contaminations or paralogs. However, if secondary sequences from multiple specimens of one cluster match to a different cluster, the likelihood that they are paralogs is higher. Note that this technique does not allow for distinguishing between low-level contamination and paralogs.

8. 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 [51]. The number of species was estimated using three different thresholds (2%, 3%, 4%). These widely used thresholds were empirically validated for Phoridae by downloading 79,506 phorids COI sequences to from NCBI using the following search terms: txid36164[Organism] AND (COI[Gene Name] OR Cox1[Gene Name] OR COXI[Gene Name] OR “cytochrome oxidase subunit 1”[Gene Name] OR “cytochrome c oxidase subunit 1”[Gene Name] OR “cytochrome oxidase subunit I”[Gene Name]). The data were filtered to exclude unidentified barcode records (containing sp. or

identified as Phoridae BOLD) and sequences that had <500 bp overlap with the 658 bp COI 5' barcoding region. Four additional sequences were excluded as they resulted in a frameshift in the sequence alignment conducted using MAFFT. The final alignment contained 12,072 DNA barcodes corresponding to 106 species, 74 of which had multiple barcodes. The barcodes corresponding to these 74 species were used to assess the distribution of pairwise intraspecific genetic distances. All 12,072 barcodes were clustered at various distance thresholds (1-10%) and the congruence with morphology was assessed using match ratios. Overall the 3% mOTUs were used for estimating species richness. Here we used EstimateS9 [53], and changed the diversity settings to use the classical formula of Chao1. This is because coefficient of variation of the abundance or incidence distribution was >0.5.

9. 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. For the strictest evaluation, this would often require genitalia preparations, but for reasons of scope, this was here only carried out for one species. For this species we illustrate how the haplotype network obtained with the median joining method in PopART (Fig. 7) [54] guides the 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 illustrated in Fig. 6. 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 [55].

List of abbreviations:

mOTU: molecular Operational Taxonomic Units

NGS: Next Generation Sequencing

NuMTs: Nuclear mitochondrial DNA sequences

BIN: Barcode Index Number

DECLARATIONS

Ethics approval and consent to participation:

The specimens were collected with valid permits (see acknowledgements).

Consent for publication: Not applicable

Availability of data and materials

<u>Source code:</u>

- 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. Available as conda package
- 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

Others

Besides the source code, the nanopore data and input files and barcode sets will be made available upon acceptance.

Competing Interests: The authors declare 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):1590-1591.
2. Krell F-T: **Parataxonomy vs. taxonomy in biodiversity studies – pitfalls and applicability of 'morphospecies' sorting**. *Biodiversity and Conservation* 2004, **13**(4):795-812.
3. Hebert PD, Cywinska A, Ball SL, deWaard JR: **Biological identifications through DNA barcodes.** *Proceedings of Royal Society B* 2003, **270**(1512):313-321.
4. Wang WY, Srivathsan A, Foo M, Yamane SK, 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, Foo M: **$1 DNA barcodes for reconstructing complex phenomes and finding rare species in specimen-rich samples**. *Cladistics* 2016, **32**(1):100-110.
6. Yeo D, Puniamoorthy J, Ngiam RWJ, Meier R: **Towards holomorphology in entomology: rapid and cost-effective adult–larva matching using NGS barcodes**. *Systematic Entomology* 2018, **43**(4):678-691.

7. Hebert PDN, Braukmann TWA, Prosser SWJ, Ratnasingham S, deWaard JR, Ivanova NV, Janzen DH, Hallwachs W, Naik S, Sones JE *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, Golding B, Hajibabaei M: **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, O'Connor R, Carvell C, Notton DG, Vogler AP: **A validated workflow for rapid taxonomic assignment and monitoring of a national fauna of bees (Apiformes) using high throughput barcoding**. *Molecular Ecology Resources* 2019, doi: 10.1111/1755-0998.13056.

10. Krehenwinkel H, Kennedy SR, Rueda A, Lam A, 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, Shoobridge JD, Graham N, Patel NH, Gillespie RG *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.

12. Pomerantz A, Penafiel N, Arteaga A, Bustamante L, Pichardo F, Coloma LA, Barrio-Amoros CL, Salazar-Valenzuela D, Prost S: **Real-time DNA barcoding in a rainforest using nanopore sequencing: opportunities for rapid biodiversity assessments and local capacity building.** *GigaScience* 2018, **7**(4):giy033.

13. Blanco M, Greene LK, Williams RC, Andrianandrasana L, Yoder AD, Larsen PA: **Next-generation in situ conservation and educational outreach in Madagascar using a mobile genetics lab**. *BioRxiv* 2019, doi: 10.1101/650614.

14. Menegon M, Cantaloni C, Rodriguez-Prieto A, Centomo C, Abdelfattah A, Rossato M, Bernardi M, Xumerle L, Loader S, Delledonne M: **On site DNA barcoding by nanopore sequencing**. *PLoS One* 2017, **12**(10):e0184741.

15. Maestri S, Cosentino E, Paterno M, Freitag H, Garces JM, Marcolungo L, Alfano M, Njunjic I, Schilthuizen M, Slik F *et al*: **A rapid and accurate MinION-based workflow for tracking species biodiversity in the field**. *Genes* 2019, **10**(6):468.

16. Wick RR, Judd LM, Holt KE: **Performance of neural network basecalling tools for Oxford Nanopore sequencing**. *Genome biology* 2019, **20**:129.

17. Srivathsan A, Bilgenur B, Wang W, Tan, WX, Bertrand D, Ng AHQ, Boey EJH, Koh JJY, Nagarajan N, Meier R: **A MinION™-based pipeline for fast and cost-effective DNA barcoding.** *Molecular Ecology Resources* 2018, **18**(5):1035-1049.

18. **Phorid Catalog** [http://phorid.net/pcat/]

19. Brown BV, 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.

20. Stork NE: **How Many Species of Insects and Other Terrestrial Arthropods Are There on Earth?** *Annu Rev Entomol* 2018, **63**:31-45.

21. Zhang ZQ: **Animal biodiversity: An introduction to higher-level classification and taxnomic richness**, vol. 3148: Magnolia Press; 2011.

22. Forbes AA, Bagley RK, Beer MA, Hippee AC, Widmayer HA: **Quantifying the unquantifiable: why Hymenoptera – not Coleoptera – is the most speciose animal order**. *BMC Ecology* 2018, **18**(21).

23. Hebert PD, Ratnasingham S, Zakharov EV, Telfer AC, Levesque-Beaudin V, Milton MA, Pedersen S, Jannetta P, deWaard JR: **Counting animal species with DNA barcodes: Canadian insects**. *Philos Trans R Soc Lond B Biol Sci* 2016, **371**(1702).

24. Borkent ART, Brown BV, Adler PH, Amorim DDS, Barber K, Bickel D, Boucher S, Brooks SE, Burger J, Burington ZL *et al*: **Remarkable fly (Diptera) diversity in a patch of Costa Rican cloud forest: Why inventory is a vital science**. *Zootaxa* 2018, **4402**(1).

25. Brown BV, Borkent A, Adler PH, De Souza Amorim D, Barber K, Bickel D, Boucher S, Brooks SE, Burger J, Burington ZL *et al*: **Comprehensive inventory of true flies (Diptera) at a tropical site**. *Communications Biology* 2018, **1**(21):8.

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

27. Vaser R, Sovic I, Nagarajan N, Sikic M: **Fast and accurate de novo genome assembly from long uncorrected reads**. *Genome research* 2017, **27**(5):737-746.

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

29. Hartop EA, 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.

30. Hartop EA, Brown BV, 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**:451-484.

31. Hartop EA, Brown BV, 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)**. In: *Biodiversity Data Journal.* vol. 4; 2016: e7756.

32. Riedel A, Sagata K, Surbakti S, Rene T, Michael B: **One hundred and one new species of Trigonopterus weevils from New Guinea**. *Zookeys* 2013(280):1-150.

33. Meier R, Zhang G, Ali F: **The use of mean instead of smallest interspecific distances exaggerates the size of the "barcoding gap" and leads to misidentification**. *Systematic Biology* 2008, **57**(5):809-813.

34. Erwin TL: **Tropical Forests: Their Richness in Coleoptera and Other Arthropod Species**. *The Coleopterists Bulletin* 1982, **36**(1):74-75.

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

36. **https://store.nanoporetech.com/kits-250/1d-sequencing-kit.html**

37. Townes H: **A light-weight Malaise trap**. *Entomological News* 1972, **83**:239-247.

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

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

40. Kurina O: **Description of four new species of *Zygomyia* Winnertz from Ethiopia and Uganda (Diptera: Mycetophilidae)**. *African Invertebrates* 2012, **53**(1):205-220.

41. Folmer O, Black M, Hoeh W, Lutz R, 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):294-299

42. Comai L, Howell T: **Barcode Generator**. In: *http://comailabgenomecenterucdavisedu/indexphp/Barcode_generator.* 2012.

43. Katoh K, Standley DM: **MAFFT multiple sequence alignment software version 7: improvements in performance and usability.** *Molecular Biology and Evolution* 2013, **30**(4):772-780.

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

45. Leray M, Yang JY, Meyer CP, Mills SC, Agudelo N, Ranwez V, Boehm JT, Machida RJ: **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.

46. Geller J, Meyer C, Parker M, 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):851-861.

47. Yeo D, Srivathsan A, Meier R: **Longer is not always better: Optimizing barcode length for large-scale species discovery and identification**. *bioRxiv* 2019: 10.1101/594952.

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

49. Srivathsan A, Sha JC, Vogler AP, 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):250-261.

50. Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, Salzberg SL: **Versatile and open software for comparing large genomes**. *Genome biology* 2004, **5**(2):R12.

51. Meier R, Shiyang K, Vaidya G, Ng PKL: **DNA Barcoding and Taxonomy in Diptera: A Tale of High Intraspecific Variability and Low Identification Success**. *Systematic Biology* 2006, **55**(5):715–728.

52. Ahrens D, Fujisawa T, Krammer HJ, Eberle J, Fabrizi S, Vogler AP: **Rarity and Incomplete Sampling in DNA-based Species Delimitatio**. *Systematic Biology* 2016, **65**(3):478-494.

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

54. Leigh JW, Bryant D: **PopART: Full-feature software for haplotype network construction.** *Methods Ecol Evol* 2015, **6**(9):1110-1116.

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

Figure Legends

Figure 1: Flowchart for generating MinION barcodes from experimental set-up to final barcodes. The novel steps introduced in this study are highlighted in green and the scripts available in minibarcoder for analyses are further indicated.

Figure 2: Effect of re-pooling on coverage of barcodes for both sets of specimens. Barcodes with coverage <50X were re-pooled and hence the coverage of these barcodes increases.

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. The plot shows that the consolidated barcodes have few ambiguities remaining.

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 MinION barcodes were trimmed to the 313 bp that were sequenced using Illumina. The overall optimal strategy is “Consolidated (namino=2)”. Optimal congruence values are highlighted in bold, and the standard deviations are in the brackets.

Dataset	**# compared with Illumina/ # 3% mOTUs**	**Accuracy**	**%Ns**	**# barcodes with errors/# >3% errors**	**mOTU richness deviation between MinION and Illumina barcodes**		
					2%	**3%**	**4%**
MAFFT	6330/663	99.6091 (0.38)	0.18 (0.2)	4473/31	-3 (-0.43%)	**-1 (-0.15)**	-8 (-1.3)
RACON	6330/663	99.5085 (0.48)	<0.001 (0.01)	4526/37	15 (2.08%)	5 (0.75)	**-3 (-0.48)**
MAFFT + AA (namino=1)	6146/659	99.9693 (0.19)	1.17 (0.83)	269/30	-6 (-0.85%)	-4 (-0.61)	-15 (-2.7)
MAFFT + AA (namino=2)	6146/661	99.9798 (0.15)	1.27 (0.91)	218/28	-8 (-1.15%)	-6 (-0.91)	-13 (-2.2)
MAFFT + AA (namino=3)	6286/659	99.9665 (0.17)	1.29 (0.95)	350/28	-7 (-0.1%)	-4 (-0.61)	-14 (-2.2)
RACON + AA (namino=1)	6312/655	99.9618 (0.19)	0.3 (0.44)	419/27	5 (0.72%)	**1 (0.15)**	**-3 (-0.48)**
RACON + AA (namino=2)	6312/655	99.9711 (0.17)	0.45 (0.52)	326/27	2 (0.29%)	-2 (-0.31)	-5 (-0.81)
RACON + AA (namino=3)	6312/659	99.9664 (0.17)	0.3 (0.57)	383/29	4 (0.57%)	-2 (-0.3)	-8 (-1.28)
Consolidated (namino=1)	6230/661	99.9855 (0.11)	0.38 (0.4)	191/25	-1 (-0.14%)	-2 (-0.30)	**-3 (-0.48)**
Consolidated (namino=2)	6255/661	**99.9864 (0.11)**	**0.43 (0.52)**	**185/25**	**0 (0%)**	-3 (-0.45)	-4 (-0.64)
Consolidated (namino=3)	6250/661	99.9762 (0.13)	0.43 (0.49)	329/26	1 (0.14%)	**-1 (-0.15)**	**-4 (-0.48)**

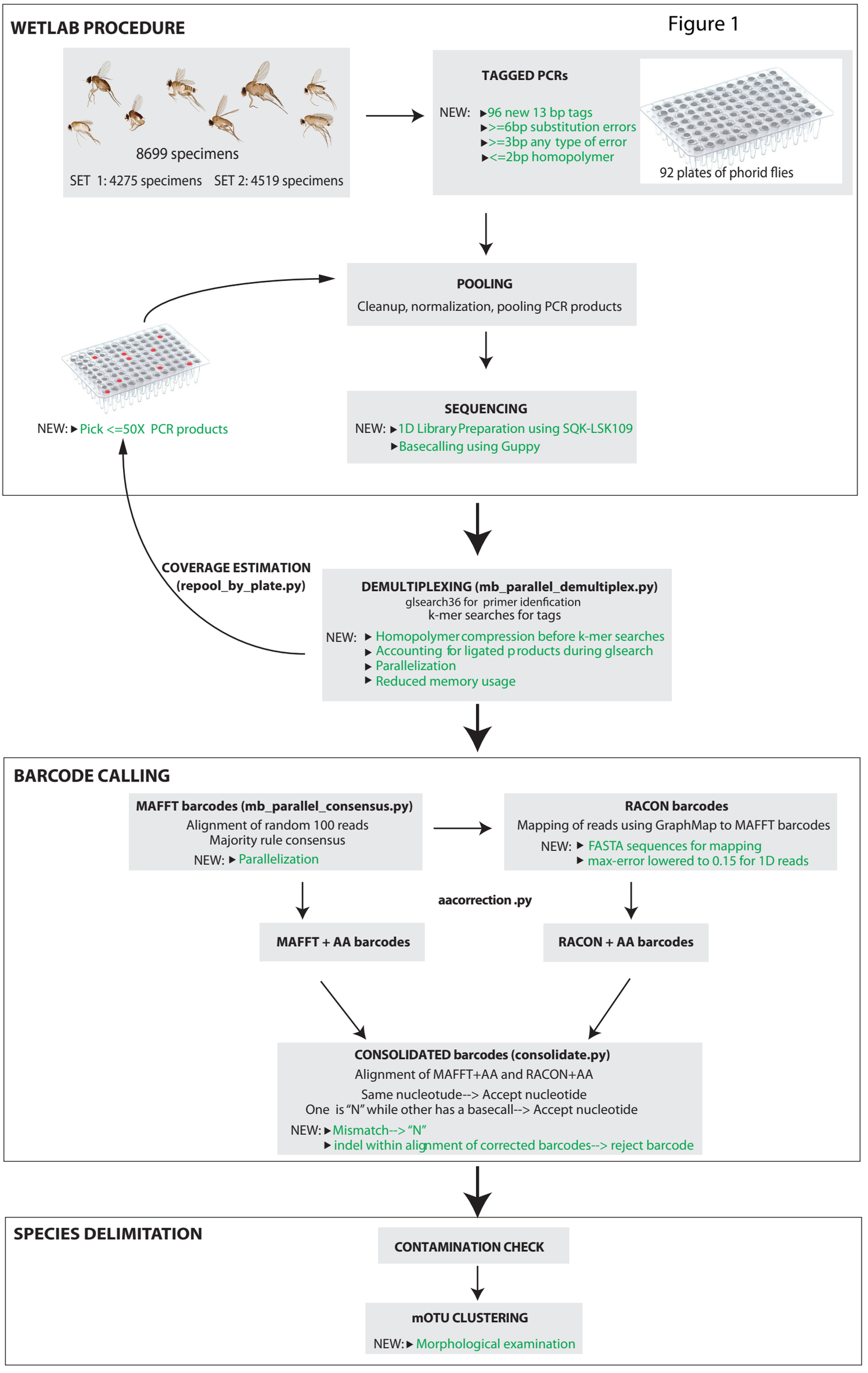
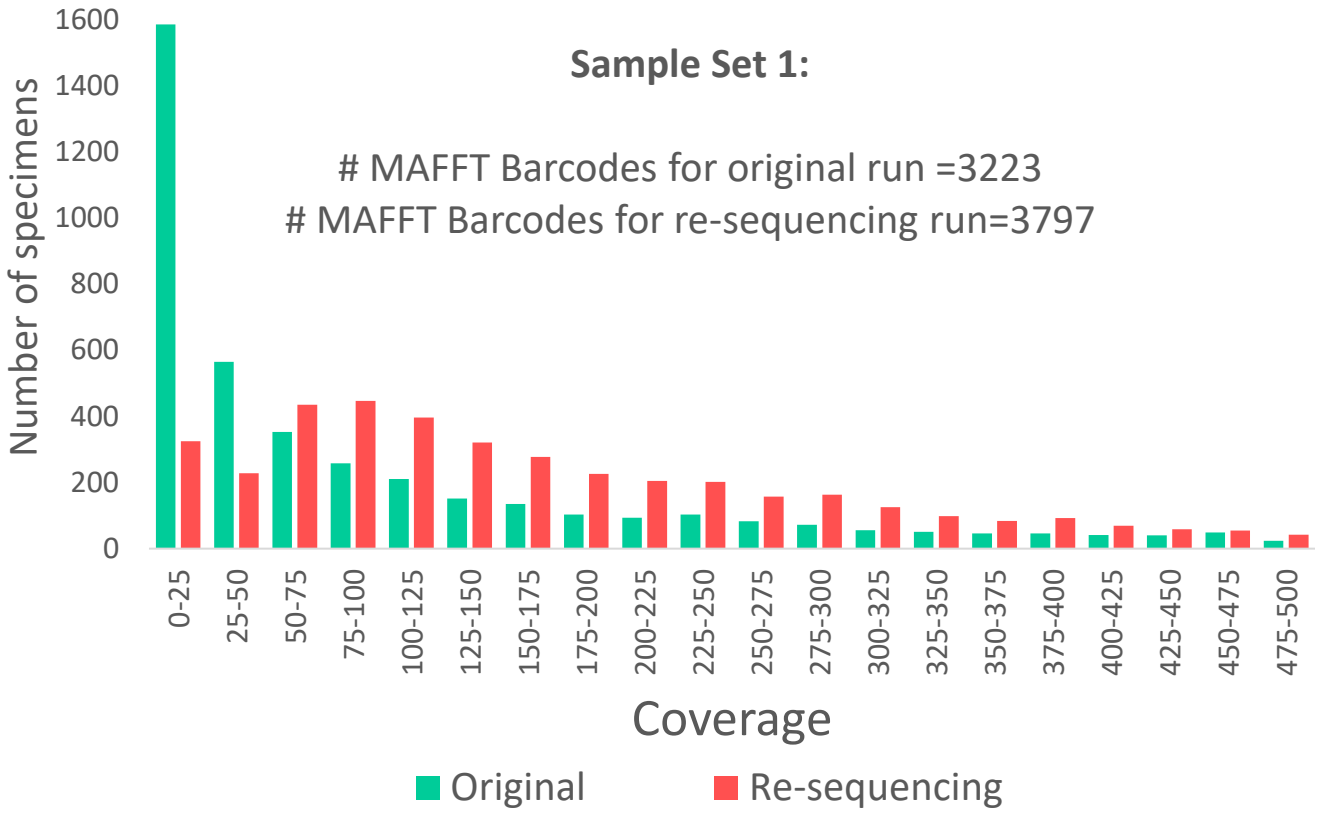

WETLAB PROCEDURE
Figure 1
8699 specimens
SET 1: 4275 specimens SET 2: 4519 specimens
TAGGED PCRs
NEW: ▸96 new 13 bp tags
▸>=6bp substitution errors
▸>=3bp any type of error
▸<=2bp homopolymer
92 plates of phorid flies
POOLING
Cleanup, normalization, pooling PCR products
SEQUENCING
NEW: ▸1D Library Preparation using SQK-LSK109
▸Basecalling using Guppy
NEW: ▸Pick <=50X PCR products
COVERAGE ESTIMATION
(repool_by_plate.py)
DEMULTIPLEXING (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

Figure 2


Sample Set 1:
MAFFT Barcodes for original run =3223
MAFFT Barcodes for re-sequencing run=3797
Number of specimens
1600
1400
1200
1000
800
600
400
200
0
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

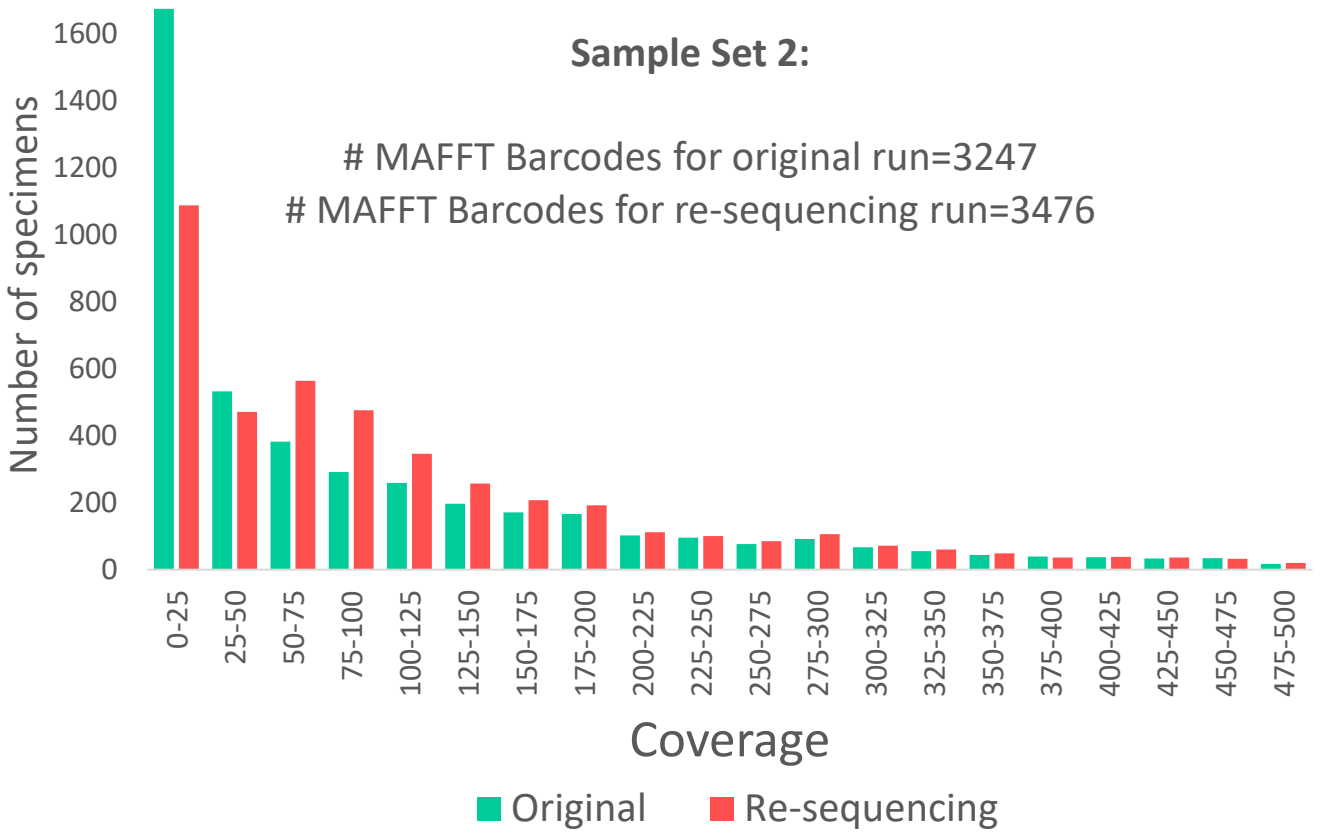

Sample Set 2:
MAFFT Barcodes for original run=3247
MAFFT Barcodes for re-sequencing run=3476
Number of specimens
1600
1400
1200
1000
800
600
400
200
0
0-25
25-50
50-75
75-100
100-125
125-150
150-175
175-200
200-225
225-250
250-275
275-300
300-325
325-350
350-375
375-400
400-425
425-450
450-475
475-500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

Figure 3

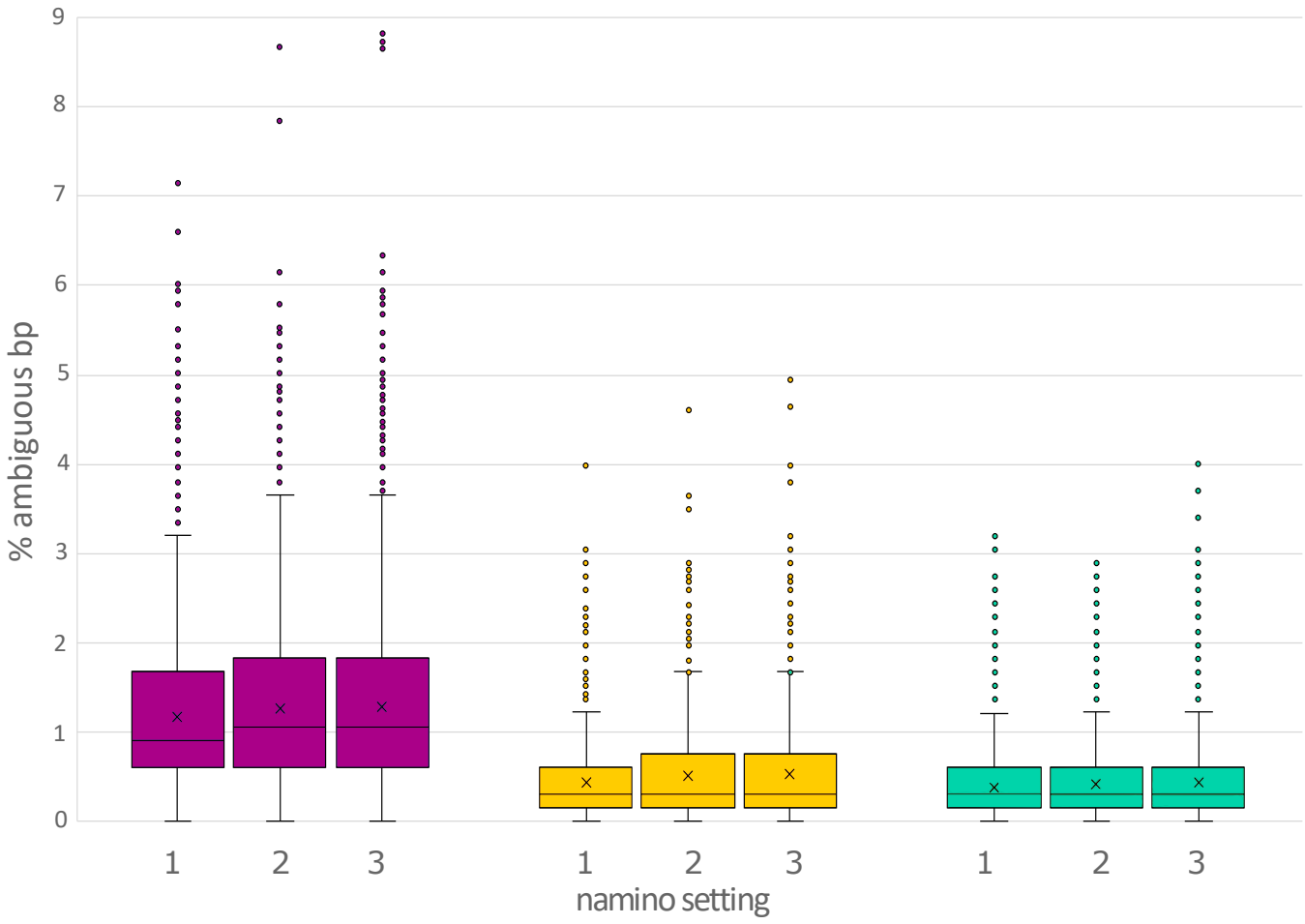


Figure 4

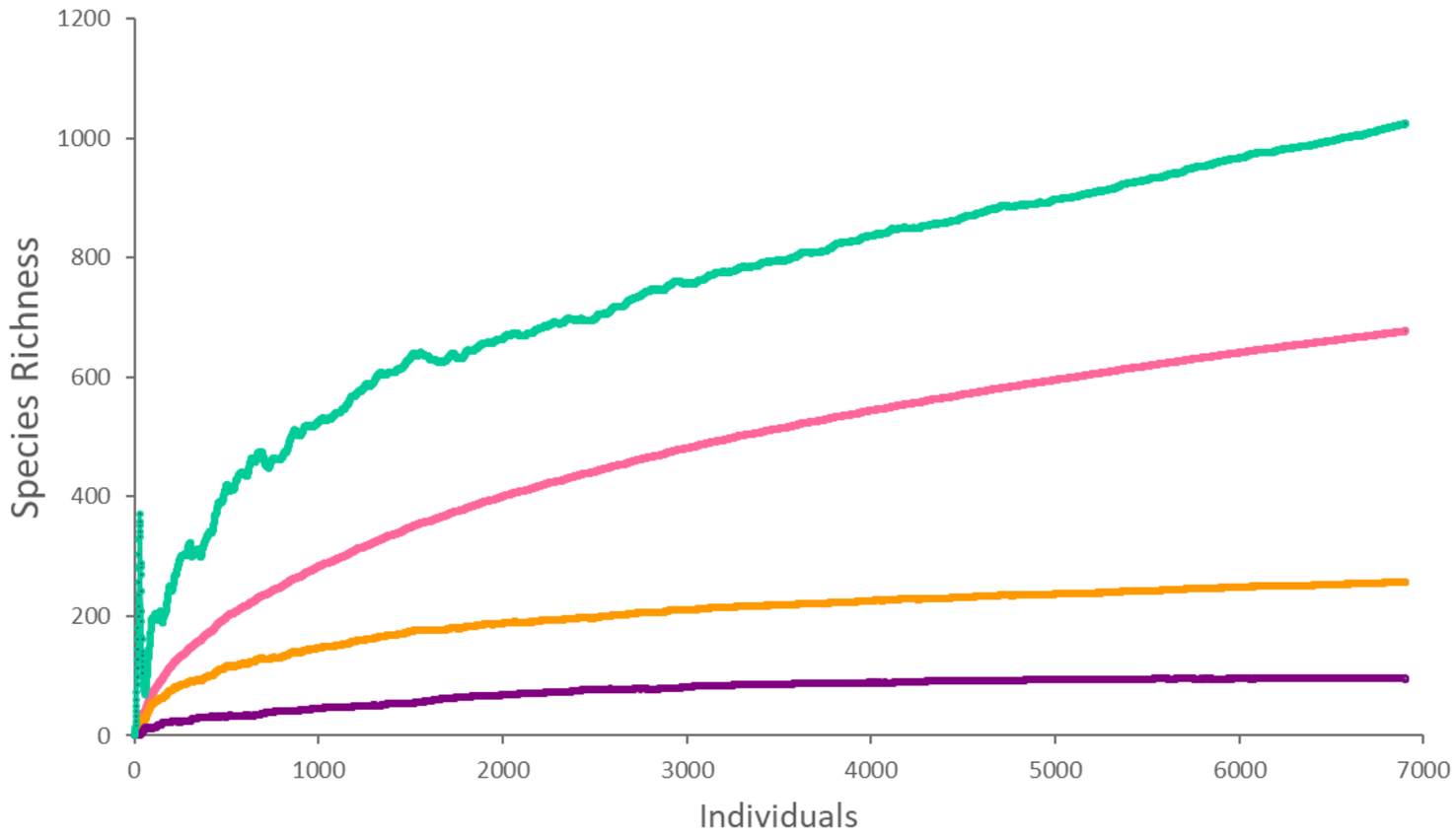
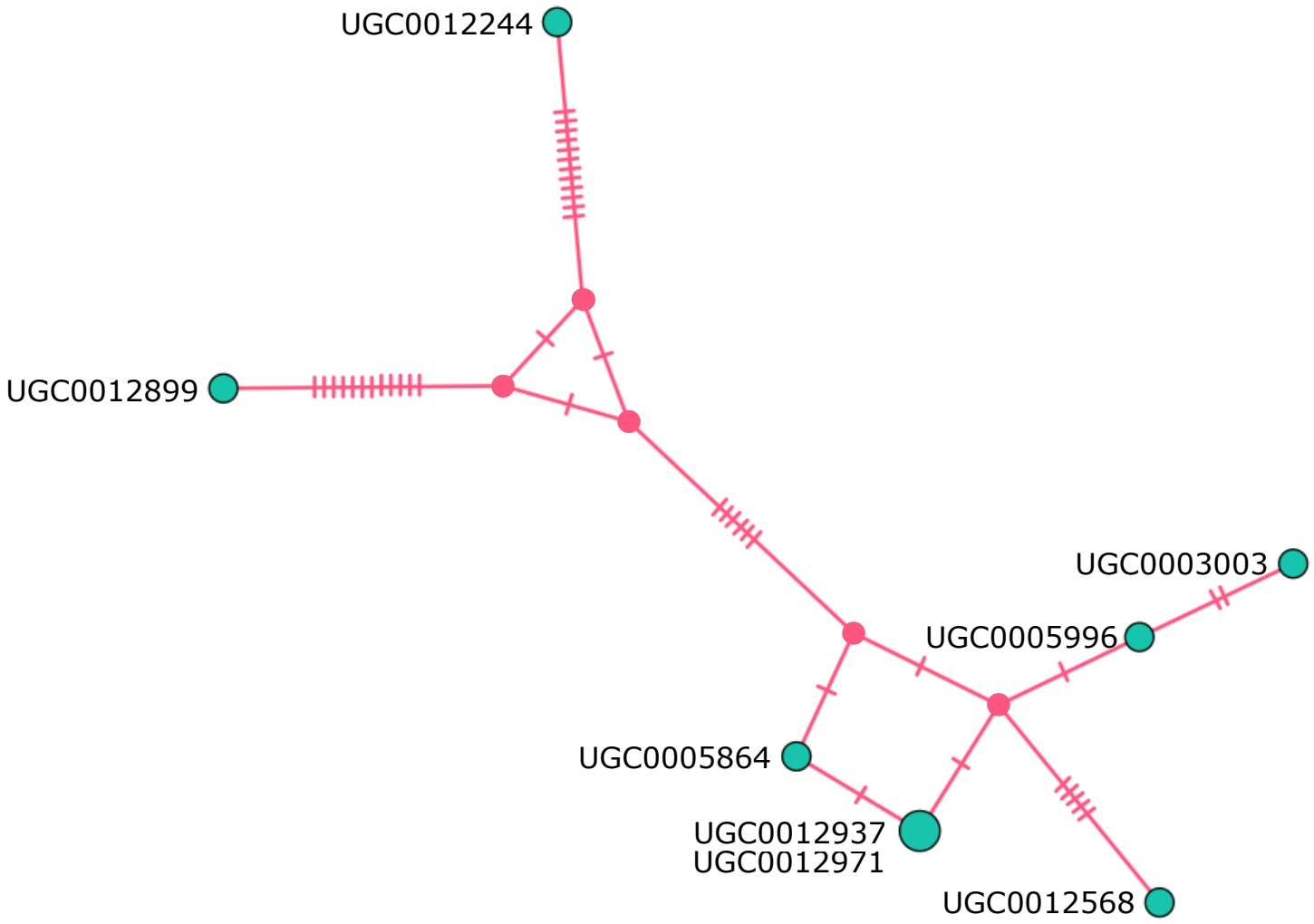

Species Richness
1200
1000
800
600
400
200
0
0
1000
2000
3000
4000
5000
6000
7000
Individuals

Figure 5

Figure 6

Figure 7


UGC0012244
UGC0012899
UGC0003003
UGC0005996
UGC0005864
UGC0012937
UGC0012971
UGC0012568