MinION barcodes: biodiversity discovery and identification by everyone, for everyone

Amrita Srivathsan[1], Leshon Lee[1], Kazutaka Katoh[2,3], Emily Hartop[4,5], Sujatha Narayanan Kutty[1,6], Johnathan Wong[1], Darren Yeo[1], Rudolf Meier[1,7]

[1] Department of Biological Sciences, National University of Singapore, Singapore

[2] Research Institute for Microbial Diseases, Osaka University, Japan

[3] Artificial Intelligence Research Center, AIST, Tokyo, Japan

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

[5] Station Linné, Öland, Sweden

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

[7] Museum für Naturkunde, Leibniz Institute for Evolution and Biodiversity Science, Center for Integrative Biodiversity Discovery, Berlin, Germany

Abstract

Background

DNA barcodes are a useful tool for discovering, understanding, and monitoring biodiversity which are critical tasks at a time of rapid biodiversity loss. However, widespread adoption of barcodes requires cost-effective and simple barcoding methods. We here present a workflow that satisfies these conditions. It was developed via “innovation through subtraction” and thus requires minimal lab equipment, can be learned within days, reduces the barcode sequencing cost to <10 cents, and allows fast turnaround from specimen to sequence by using the portable, real-time sequencer MinION.

Results

We describe cost-effective and rapid procedures for barcoding individual specimens with MinION sequencing. We illustrate how tagged amplicons can be obtained and sequenced with the portable, real-time MinION sequencer in many settings (field stations, biodiversity labs, citizen science labs, schools). We also provide amplicon coverage recommendations that are based on several runs of the latest generation of MinION flow cells (“R10.3”) which suggest that each run can generate barcodes for >10,000 specimens. Next, we present a novel software, ONTbarcoder, which overcomes the bioinformatics challenges posed by MinION reads. The software is compatible with Windows 10, Macintosh, and Linux, has a graphical user interface (GUI), and can generate thousands of barcodes on a standard laptop within hours based on only two input files (FASTQ, demultiplexing file). We document that MinION barcodes are virtually identical to Sanger and Illumina barcodes for the same specimens (>99.99%) and provide evidence that MinION flow cells and reads have improved rapidly since 2018.

Conclusions

We propose that barcoding with MinION is the way forward for government agencies, universities, museums, and schools because it combines low consumable and capital cost with scalability. Small projects can use the flow cell dongle (“Flongle”) while large projects

can rely on MinION flow cells that can be stopped and re-used after collecting sufficient data for a given project.

1. Background

DNA sequences have been used for identification and taxonomic purposes for decades [1-3], but for most of this time they have been akin to mobile phones in the 1990s: of limited value due to sparse signal coverage and high cost. Obtaining barcodes was too complicated and expensive despite the development of effective DNA extraction protocols [4], fast Sanger sequencing protocol [5], and the establishment of the Canadian Centre for DNA Barcoding (CCDB), which is the primary sequencing facility for the International Barcode of Life Consortium (iBOL: https://ibol.org/resources/sequencing-facility/). After 15 years and the investment of >200 million USD in facilities and barcoding, ca. 8 million animal barcodes are available for searches in the database "BOLD Systems" of which ca. 6 million sequences can be downloaded as "public data" (April 2021: [6]). Combined with barcodes from NCBI, they are now a valuable resource for the global biodiversity community.

However, the cost of barcodes has remained high (http://ccdb.ca/pricing/) and the most widely used approach to large-scale barcoding still relies on sending specimens from all over the world to the CCDB in Guelph (Canada). For example, >85% of all arthropod barcodes in BOLD Systems were generated by the CCDB with more than >60% of the voucher specimens remaining in the centre (https://www.boldsystems.org/index.php/Taxbrowser_Taxonpage?taxid=20). Unfortunately, this model for barcoding interferes with real-time biodiversity monitoring and specimen accessibility. We therefore here argue that barcoding has to be decentralized. We show that this can be achieved through "innovation through subtraction," which yields simplified and cost-effective solutions for generating barcode amplicons in molecular laboratories with very basic equipment. Combined with the use of MinION sequencers, these simplifications allow for generating barcodes almost anywhere by biologists and citizen scientists alike.

A decentralized model for monitoring the world's biodiversity is necessary given the scale, urgency, and importance of the task at hand. For example, even if there were only 10 million

species of metazoan animals on the planet [7] and a new species was discovered with every 50th specimen that is processed, species discovery with barcodes will require the sequencing of 500 million specimens [8]. Yet, species discovery is only a small part of the biodiversity challenge in the 21st century. Biodiversity loss is now considered by the World Economic Forum as one of the top three global risks based on likelihood and impact for the next 10 years [9] and Swiss Re estimates that 20% of all countries face ecosystem collapse as biodiversity declines [10]. This implies that biodiversity discovery and monitoring will require in the future completely different scales than in the past. The old approaches thus need rethinking because all countries will need real-time distributional and abundance information in order to develop effective conservation strategies and policies. In addition, they will need information on how species interact with each other and the environment [11].

Barcodes were initially intended as an identification tool for biologists [1]. Thus, most projects focused on taxa with a large following in biology (e.g., birds, fish, butterflies) [12] Yet, despite targeting taxa with well-understood diversity, the projects struggled with barcoding >75% of the described species in these groups [12]. When the pilot barcoding projects ran out of tissues from identified specimens, they started targeting unidentified specimens; i.e., DNA barcoding morphed into a technique that was used for biodiversity discovery (“dark taxa”: [12, 13]). This shift towards biodiversity discovery was gradual and incomplete because the projects used a “hybrid approach” that started with sorting specimens to “morphospecies” before barcoding one or a few specimens for each morphospecies/sample (e.g., [14-20]). However, this approach is not ideal for biodiversity discovery and monitoring because morphospecies sorting is labour-intensive and of unpredictable quality because it is heavily dependent on the taxonomic expertise of the sorters [21, 22]. This is why it is preferable to reverse the workflow [23] by barcoding all specimens first and assessing congruence with morphology afterwards [24, 25]. This approach has the additional benefit that it yields quantitative data and corroborated species-

level units. However, it requires efficient and low-cost barcoding methods that are also suitable for biodiverse countries with limited science funding.

Fortunately, such cost-effective barcoding methods are now becoming available. This is partially due to the replacement of Sanger sequencing with second- and third-generation sequencing technologies that have lower sequencing costs [23, 26-33]. These changes mean that the widespread application of the reverse workflow is now feasible for tackling the species-level diversity of those metazoan clades that are so specimen- and species-rich that they have been neglected in the past [33, 34]. Many of these clades have high spatial species turnover, requiring many localities in each country to be sampled and large numbers of specimens to be processed [32]. Such intensive processing is best achieved close to the collecting locality to avoid the delays and costs of shipping samples across continents. This is now feasible because biodiversity discovery can be readily pursued in decentralized facilities at varied scales. Indeed, accelerated biodiversity discovery is a rare example of a big science initiative that allows for meaningful engagement of students and citizen scientists, which can enhance significantly biodiversity education and appreciation [35-38]. This is especially so when stakeholders not only barcode, but also image specimens, determine species abundances, and map distributions of newly discovered species. All of which can be based on specimens collected in the citizens' own backyard.

But can such decentralized biodiversity discovery really be effective? Within the last five years, students and interns in the laboratory of the corresponding author at the National University of Singapore barcoded >330,000 specimens. After analyzing the first >140,000 barcoded specimens for selected taxa representing different ecological guilds, the alpha and beta diversity of Singapore's arthropod fauna was analyzed based on ~8,000 putative species which revealed that some habitats were unexpectedly species-rich and harboured unique faunas (e.g., mangroves, freshwater swamp: [32, 39]). Barcodes even helped with the conservation of charismatic taxa when they were used to identify the larval habitats for

more than half of Singapore's damsel- and dragonfly species [40] and facilitated species interaction research and biodiversity surveys based on eDNA [41, 42]. Biodiversity appreciation by the public was fostered by featuring newly discovered species and their species interactions on "Biodiversity of Singapore" (BOS >15,000 species: https://singapore.biodiversity.online/), dozens of new species have been described, and the descriptions of another 150 species are being finalized [43-51].

Barcoding metazoan specimens require the successful completion of three steps: (1) obtaining DNA template, (2) amplifying *COI* via PCR, and (3) sequencing the *COI* amplicon. Many biologists learn these techniques in university for a range of different genes – from those that are easy to amplify (short fragments of ribosomal and mitochondrial genes with well-established primers) to those are difficult (long, single-copy nuclear genes with few known primers). Fortunately, amplification of short mitochondrial markers like *COI* does not require the same level of care as nuclear markers. Learning how to barcode efficiently is hence an exercise of unlearning unnecessary procedures. Note that this unlearning comes with cost savings which are particularly vital for boosting biodiversity research where it is most needed: in biodiverse countries with limited science funding. Our manuscript therefore has a more comprehensive Methods section than most publications, because we do not only describe which methods we used for our experiments, but also why certain alternative methods were avoided. In addition, we provide videos that illustrate the techniques (https://www.youtube.com/channel/UC1WowokomhQJRc71FmsUAcg).

Simplified laboratory techniques for obtaining barcode amplicons are important, but they need to be complemented with efficient and cost-effective sequencing techniques. Therefore, we here also test the capabilities of the latest flow cells used in Oxford Nanopore Technologies (ONT) instruments. They have the advantage of being inexpensive and yielding data quickly by passing single-stranded DNA through a nanopore and using the current fluctuations for reconstructing the DNA sequence [52]. ONT's MinION is especially

suitable for decentralized barcoding because it is small and inexpensive. However, its use for barcoding has been unpopular because of low sequence read accuracy (85% - 95%: [52, 53]). This meant that the data had to be analyzed with complex bioinformatics pipelines that are not suitable for widespread use.

Recently, three significant changes occurred. Firstly, ONT released a low-cost and capacity flow cell (Flongle) that only has 126 pores (126 channels) instead of the customary 2048 (512 channels) of a full MinION flow cell. We here test whether Flongle is a promising tool for small barcoding projects that need quick turnaround times for a few hundred specimens. Secondly, ONT released a new flow cell chemistry for full flow cells ("R10.3") where the nanopores have a dual instead of a single reader-head. Dual reading has altered the read error profile by giving better resolution to homopolymers and improved consensus accuracy [54, 55]. Lastly, ONT released high accuracy (HAC) basecalling (https://nanoporetech.com/about-us/news/new-research-algorithms-yield-accuracy-gains-nanopore-sequencing). All three innovations are here tested using six different amplicon pools of very different sizes (191 - 9,932 specimens). Overall, we find that the innovations were very effective at improving read quality and quantity. This meant that we could develop and release a new bioinformatics software, "ONTbarcoder," which is fast and user-friendly in that it has a GUI, is compatible with all major operating systems, and does not require the installation of third-party software.

Results

Performance of new flow cells and high-accuracy basecalling

We used six pools of amplicons to test the new ONT products. The pools contained amplicons for 191 - 9,932 specimens and were run for 15-49 hours (Table 1). The amplicons were tagged with a 13-bp index to facilitate demultiplexing reads to specimen specific bins. Basecalling the fast5 files using Guppy in MinIT under the high accuracy (HAC) model was still very slow and took 12 days for the largest dataset (*Palaearctic Phoridae (658 bp)* (Table

1). However, it yielded good quality reads that could be demultiplexed at high rates for the four R10.3 MinION datasets (= 30-49%). The only exception was the *Palaearctic Phoridae (313 bp)* dataset (15.5%). Flongle datasets showed overall also lower demultiplexing rates (17-21%).

Table 1. Datasets generated in this study and the results of barcoding using ONTbarcoder at 200X coverage (Consensus by Length) and 100X coverage (Consensus by Similarity).

Dataset Name	**Flow cell details Run time/Guppy version**	**Raw reads/reads passing length threshold/reads of suitable length/ demultiplexed**	**Demultiplexing rate/# QC compliant barcodes /# Filtered barcodes with 1N/# Filtered barcodes with >1N /# Unreliable barcodes**
MinION R10.3 Datasets			
Mixed Diptera (658 bp, N=511)	R10.3: reused flow cell: 71 pores according to QC, but 500+ active during run Runtime: 27.5 hrs Guppy: 4.2.3+f90bd04	3,864,000/3,425,357/3,560,389/1,544,758	43.39%/495/2/5/8 Total success rate= 502/511 (98.2%)
Afrotropical Phoridae (658 bp, N=4,275)	R10.3: new flow cell: QC: 1,101 pores Runtime: 49.5 hrs Guppy: 4.0.11+f1071ce	6,838,903/5,465,164/5,474,306/2,681,029	48.97%/3,722/121/59/247 Total success rate= 3905/4275 (91.3%)
Palaearctic Phoridae (658 bp, N=9,932)	R10.3: new flow cell: QC: 1,239 pores Runtime: 47.5 hrs Guppy: 4.2.3+f90bd04	16,595,984/15,658,174/16,100,505/5,012,489	31.13%/8,026/108/231/780 Total success rate= 8,365/9,932 (84.2%)
Palaearctic Phoridae (313 bp, N=9,929)	R10.3: new flow cell: QC: 1,297 pores Runtime: 37 hrs Guppy: 4.2.3+f90bd04	13,690,869/13,221,764/10,366,455/12,983,260/2,015,135	15.52%/8,705/118/112/899 Total success rate= 8,935/9,929 (90%)
Flongle Datasets			
Mixed Diptera Subsample (658 bp, N=257)	Flongle: new QC: 81 pores Runtime: 24 hrs Guppy: v 4.0.11+f1071ce	294,896/222,189/190,952/33,270	17.42%/185/35/20/9 Total success rate= 240/257 (93.4%)
Chironomidae (313 bp, N=191)	Flongle: new QC: 74 pores Runtime: 15 hrs Guppy: 4.2.3+f90bd04	560,062/525,087/504,621/108,574	21.52%/178/1/2/6 Total success rate= 181/191 (94.8%)

We then investigated barcode accuracy (Figure 1) by directly aligning the MinION barcodes with Sanger and Illumina barcodes for the same specimens. We find that MinION barcodes are virtually identical (>99.99% identity, Table 2). We furthermore established that the number of ambiguous bases ("N") is very low for barcodes obtained with R10.3 (<0.01%). Indeed, more than 90% of all barcodes are entirely free of ambiguous bases. In comparison, Flongle barcodes have a slightly higher proportion of ambiguous bases (<0.06%). They are concentrated in ~20% of all sequences so that 80% of all barcodes again lack Ns. This means that MinION barcodes well exceed the Consortium for the Barcode of Life (CBOL) criteria for "barcode" designation with regard to length, accuracy, and ambiguity (Ratnasingham and Hebert 2013).

Table 2. Quality assessment of barcodes generated by ONTbarcoder at 200X read coverage (Consensus by Length) and 100X coverage (Consensus by Similarity). The accuracy of MinION barcodes is compared with the barcodes obtained for the same specimens using Illumina/Sanger sequencing. Errors are defined as sum of substitution or indel errors. Denominators are the total number of nucleotides assessed.

Dataset	**No. of comparison barcodes**	**No. of barcodes with errors/No. of errors/% identity**	**# of Ns/%Ns**
R10.3: Mixed Diptera: Sanger barcodes available	476	2/10/99.997%	19 (0.006%)
R10.3: Afrotropical Phoridae: Illumina barcodes available*	3316	23/48/99.995%	284 (0.011%)
Flongle-Mixed Diptera Subsample: Sanger barcodes available	231	5/8/99.994%	91 (0.058%)

*5 barcodes with very high distances from reference were excluded for R10.3: Afrotropical Phoridae dataset as they likely represent lab contamination (see Srivathsan, Hartop et al. (2019)

Rarefaction at different read coverage levels reveals that 80-90% of high-quality barcodes are obtained within a few hours of sequencing. In addition, the number of barcodes generated by MinION exceeded or was comparable to what could be obtained with Sanger or Illumina (Figure 1). We also determined the coverage needed for obtaining reliable barcodes. For this purpose, we plotted the number of barcodes obtained against coverage

(Figure 2). This revealed that the vast majority of specimens yield high-quality barcodes at coverages between 25x and 50x when R10.3 reads are used. Increasing coverage beyond 50x only leads to modest improvements in quality and few additional barcodes. The coverage needed for obtaining Flongle barcodes was somewhat higher, but the main difference between the R9.4 technology of the Flongle flow cell and R10.3 of the MinION flow cell was that more barcodes retained ambiguous bases even at high coverage for R9.4 data. The differences in read quality between R9.4 and R10.3 became even more obvious when the read bins for the “Mixed Diptera Subsample” were analyzed based an identical numbers of R10.3 and R9.4 reads. The barcodes based on Flongle and R10.3 data are compatible, but the R10.3 barcodes are ambiguity-free while some of the corresponding Flongle barcodes retain 1-2 ambiguous bases.

Overall, these results imply that 100x raw read coverage is sufficient for obtaining barcodes with either R10.3 or R9.4 flow cells. Given that most MinION flow cells yield >10 million reads of an appropriate length, one could, in principle, obtain 100,000 barcodes in one flow cell. However, this would require that all amplicons are represented by similar numbers of copies and that all reads could be correctly demultiplexed. In reality, only 30-50% of the reads can be demultiplexed and the number of reads per amplicon fluctuates widely (Figure 3). Very-low coverage bins tend to yield no barcodes or barcodes of lower quality (errors or Ns). These low-coverage barcodes can be improved by collecting more data, but this comes at a high cost and increased risk of a small number of contaminant reads yielding barcodes. For example, we observed that some “negative” PCR controls yielded low-quality barcodes for 4 of 106 negatives in the Palaearctic Phoridae (313 bp) and 1 of 105 negatives in the Palaearctic Phoridae (658 bp) datasets.

To facilitate the planning of barcode projects, we illustrate the trade-offs between barcode yield, sequencing time, and amount of raw data needed for six amplicon pools (Figure 4: 191-9,932 specimens). These standard curves can be used to roughly estimate the number

of raw reads needed to achieve a specific goal for a barcoding project of a specific size (e.g., obtaining 80% of all barcodes for a project with 1000 amplicons). Note that the number of raw reads is displayed in real-time in MinKNOW so that the run can be terminated when the target number of reads has been reached. The number of recoverable barcodes in Figure 4 was set to the number of all error-free, filtered barcodes obtained in an analysis of all data. We would argue that this is a realistic estimate of recoverable barcodes given the saturation plots in Figure 1 that suggest that most barcodes with significant amounts of data have been called at 200x coverage. Note, however, that Figure 4 can only provide very rough guidance on how many reads are needed because, for example, the demultiplexing rates differ between flow cells and different amplicon pools have very different read abundance distributions (see Figure 3).

We further compared the barcodes obtained with ONTbarcoder with those obtained via the recent software NGSpeciesID [56]. NGSpeciesID often provides multiple consensus barcodes for the same set of reads for a specimen. We here only compare the consensus barcodes supported by the highest number of reads. When compared to Illumina or Sanger reference barcodes, the barcodes obtained via ONTbarcoder have fewer errors (Table 3: 25-118 erroneous barcodes for NGSpeciesID vs 5-28 erroneous barcodes for ONTbarcoder). In addition, NGSpeciesID also yielded consensus barcodes for negative controls because it performs no quality control based on length, translation or other criteria. Most of the corresponding read sets yielded no barcodes with ONTbarcoder due to rigorous quality checks and low read count filters. For example, NGSpeciesID formed 32 negative consensus barcodes for Afrotropical Phoridae 658 dataset (demultiplexed by minibar) with 30 being represented by <5 sequences.

Table 3. Comparison of barcodes obtained by ONTbarcoder with NGSpeciesID (green=highest number of correct and lowest number of errors for datasets). NGSpeciesID was applied once to the demultiplexed reads obtained with ONTbarcoder and once to those obtained with minibar. Barcode calling was done using all demultiplexed reads as well as 200X subset. ONTbarcoder was run at 200X read coverage (Consensus by Length) and

100X coverage (Consensus by Similarity). The accuracy of MinION barcodes is only compared to reference barcodes obtained with Illumina/Sanger sequencing. Errors are defined as sum of substitution and indel errors.

	ONTBarcoder 200X	ONTBarcoder dem[1] + NGSpeciesID All reads	ONTBarcoder dem[1] + NGSpeciesID 200X	minibar + NGSpeciesID All reads	minibar + NGSpeciesID 200X
	Mixed Diptera Subsample Flongle				
Identical+compatible[2] (# of reference barcodes)	**226** (240)	72 (241)	73 (241)	121 (239)	122 (239)
Incorrect, (1/2/3/4/5+) errors	**5 (3/1/1/0/0)**	169 (83/60/13/6/7)	168 (88/55/13/7/5)	118 (29/21/8/4/56)	119 (31/20/8/4/56)
# reads demultiplexed/reads of length 658±50	33,270/32,978			19,511/15,935	
	Mixed Diptera R10.3				
Identical+compatible[2] (# of reference barcodes)	**473** (475)	173 (478)	353 (478)	418 (478)	453 (478)
Incorrect, with (1/2/3/4/5+) errors	**2 (0/0/1/0/1)**	305 (126/97/48/17/17)	125 (91/24/5/1/4)	60 (45/9/3/0/3)	25 (20/2/0/0/3)
# reads demultiplexed/reads of length 658±50	1,544,768/1,532,408			1,540,751/1,378,910	
	Afrotropical Phoridae R10.3				
Identical+compatible[2] (# of reference barcodes)	**3293** (3321)	2832 (3339)	3111 (3341)	3193 (3340)	3245 (3340)
Incorrect, (1/2/3/4/5+) errors	**28 (15/5/0/0/8)**	507 (362/90/19/12/24)	230 (172/31/5/3/19)	147 (90/27/3/1/26)	95 (50/13/12/1/27)
# reads demultiplexed/reads of length 658±50	2,681,029/2,668,933			3,117,597/2,782,367	

[1] dem=demultiplexing

[2] Identical barcodes match perfectly with references, and there are no ambiguities while compatible barcodes match with reference with at 100% identity but contain ambiguities. NGSpeciesID does not introduce N's in barcodes and thus these two categories were grouped for this comparison.

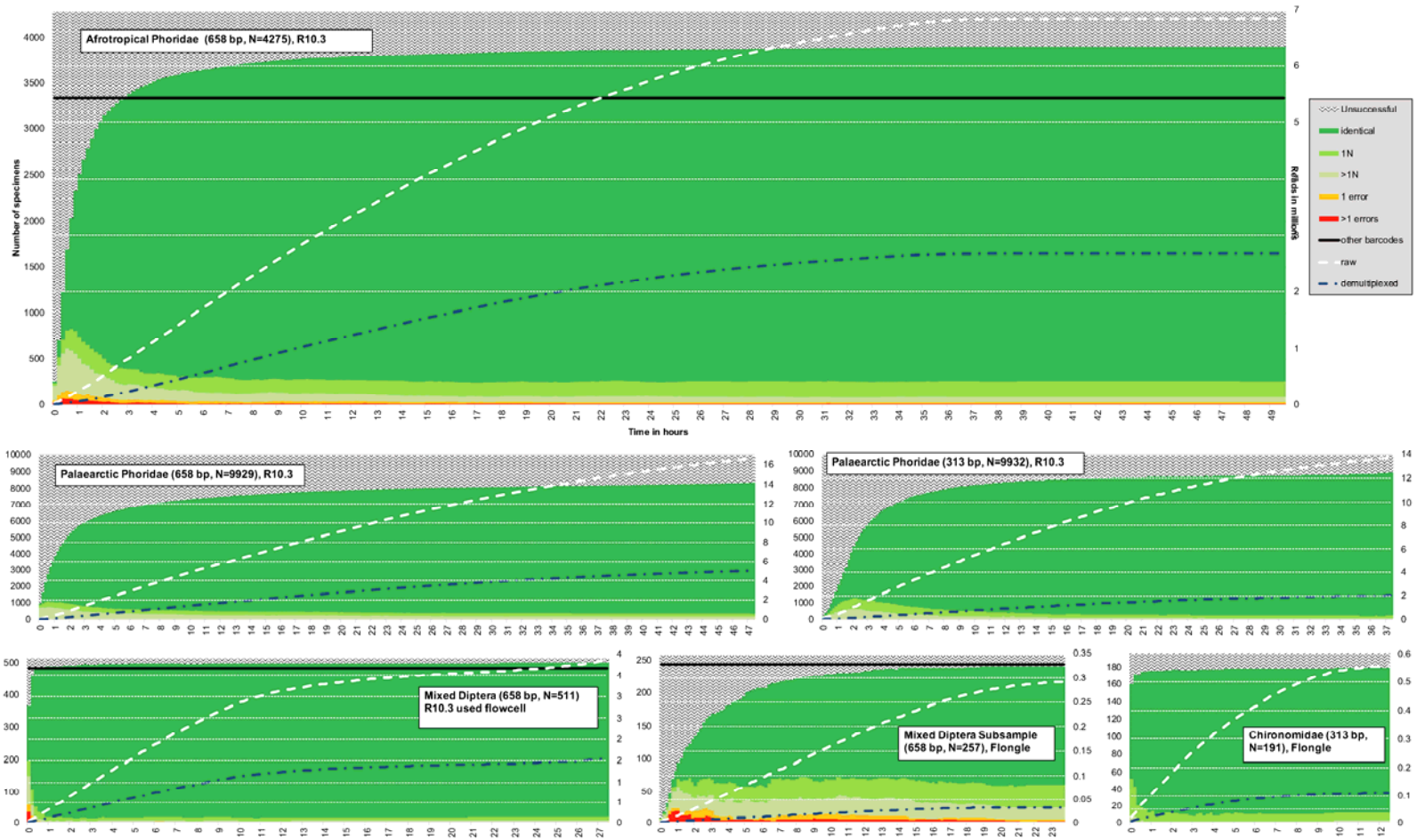

Figure 1. Rapid recovery of accurate MinION barcodes over time (in hours, x-axis) (filtered barcodes: dark green = barcodes passing all 4 QC criteria, light green = one ambiguous base; lighter green = more than 1N, no barcode = white with pattern, 1 mismatch = orange, >1 mismatch = red). The solid black line represents the number of barcodes available for comparison. White dotted line represents the amount of raw reads collected over time, blue represents number of demultiplexed reads over time (plotted against Z-axis)

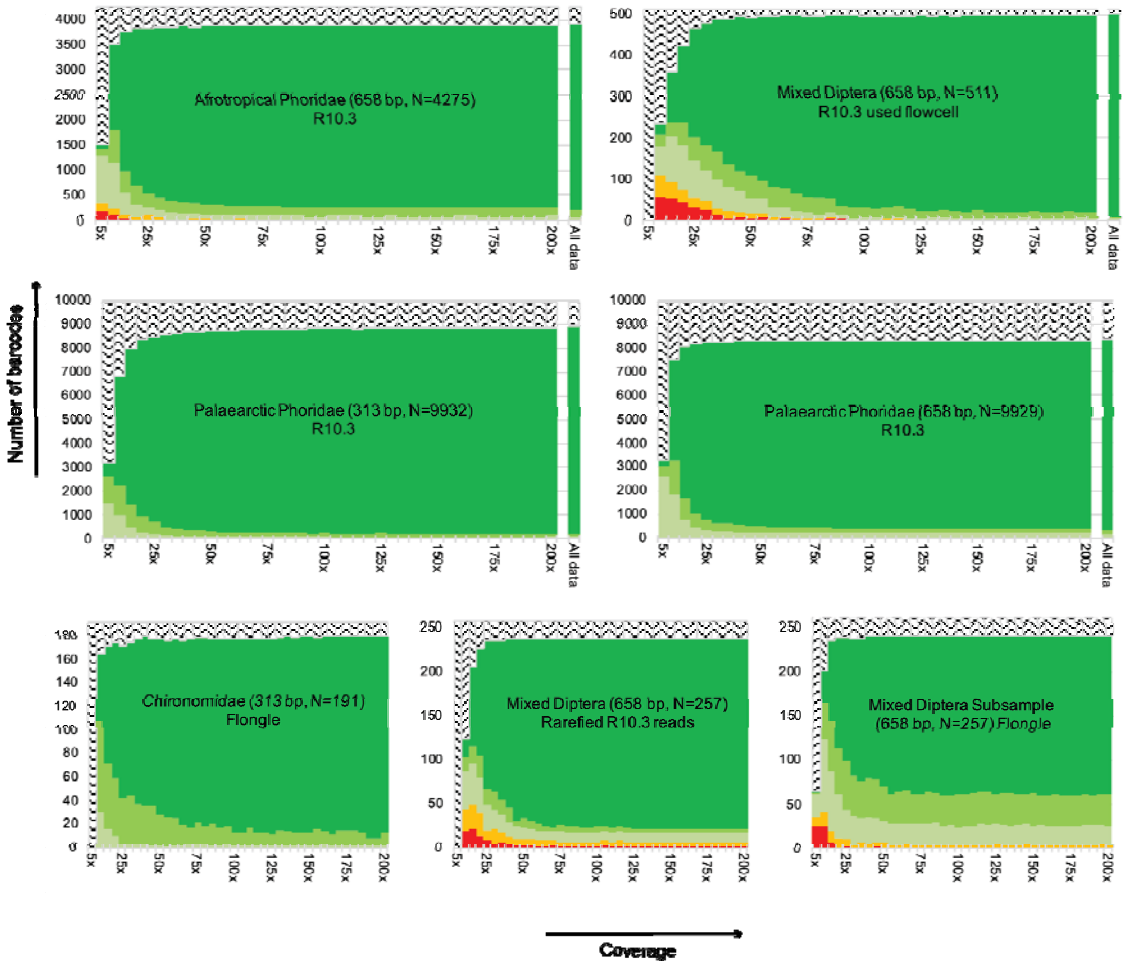

Figure 2. Relationship between barcode quality and coverage. Subsetting the data to 5-200X coverage shows that there are very minor gains to barcode quality after 25-50X coverage. (filtered barcodes: dark green = barcodes passing all 4 QC criteria, light green = one ambiguous base; lighter green = more than 1N, no barcode = white with pattern, 1 mismatch = orange, >1 mismatch = red).

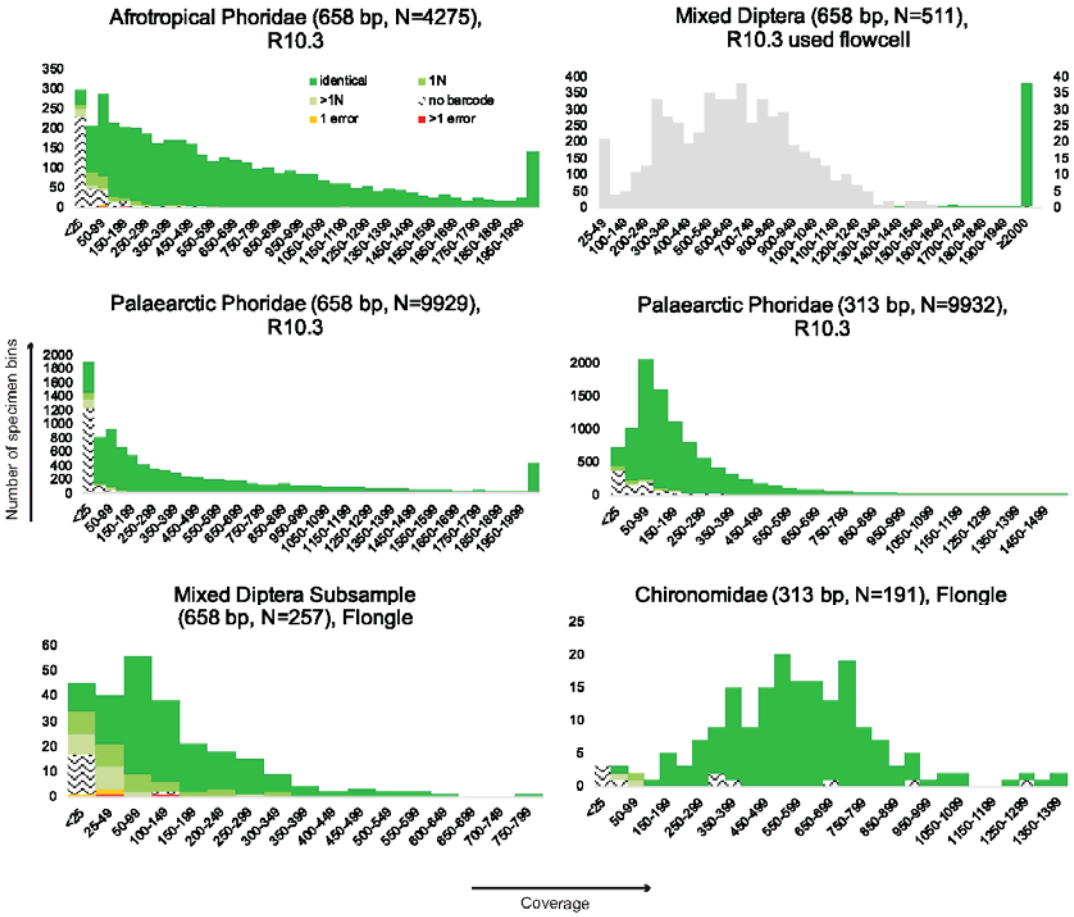

Figure 3. Read bin size distribution for six amplicon pools (color-coding as in Figs 1-2). Due to the very generous coverage for the “Mixed Diptera” dataset, we also use grey to show the bin size distribution after dividing the bin read totals by 5.

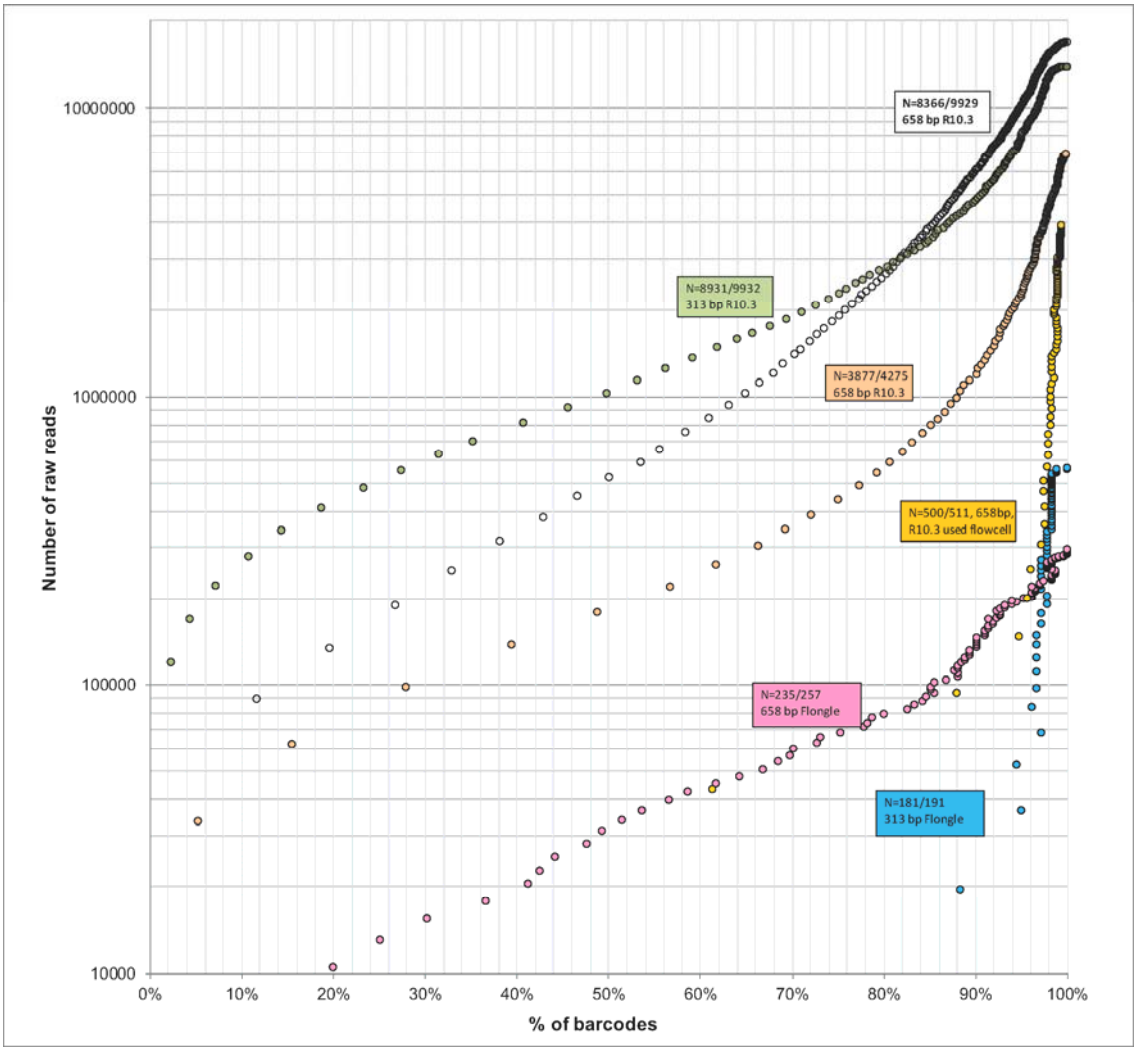

Figure 4. Relationship between barcoding success and number of raw reads for six amplicon pools (191-9932 specimens; barcoding success rates 84-97%). Percentage of barcodes recovered is relative to the final estimate based on all data.

Discussion

Democratization of Barcoding

Biodiversity research needs new scalable techniques for large-scale species discovery and monitoring. This task is particularly urgent and challenging for invertebrates that collectively make up most of the terrestrial animal biomass. We argued earlier that this is likely to be a task that requires the processing of at least 500 million specimens from all over the world with many tropical countries with limited research funding requiring much of the biodiversity

discovery work. Pre-sorting these specimens into putative species-level units with DNA sequences is a promising solution as long as obtaining and analyzing the data are sufficiently straightforward and cost-effective. We believe that the techniques described in this manuscript will help with achieving these goals.

We here show that sequencing barcode amplicons with MinION is a particularly attractive option. Firstly, MinION library preparation can be learned within hours and an automated library preparation instrument is available and eventually expected to also generate ligation-based libraries ("VolTRAX"). Secondly, MinION flow cells can accommodate projects of varying scales. Flongle can be used for amplicon pools with a few hundred products, while an R10.3 flow cell can accommodate projects with up to 10,000 specimens. The collection of data on MinION flow cells can be stopped whenever the software controlling the run ("MinKNOW") indicates that a sufficiently high number of reads have been acquired. Flow cells can then be washed and re-used again although the remaining capacity declines over time so that we have only re-used flow cells up to four times.

Traditionally, the main obstacles to using MinION for barcoding have been poor read quality and high cost. Both issues are fading into the past. The quality of MinION reads has improved to such a degree that the laptop-version of our new software "ONTbarcoder" can generate thousands of very high quality barcodes within hours. There is also no longer a need to "polish" reads or to rely on external data or algorithms. The greater ease with which MinION barcodes can be obtained is due to several factors. Firstly, each flow cell now yields a much larger numbers of reads. Secondly, R10.3 reads have a different error profile which allows for reconstructing higher-quality barcodes. Thirdly, high accuracy basecalling has improved raw read quality and thus demultiplexing rates. Lastly, we can now use parameter settings for MAFFT that are designed for MinION reads. These changes mean that even low-coverage bins can yield very accurate barcodes; i.e., both barcode quality and quantity are greatly improved.

We had previously tested MinION for barcoding in 2018 and 2019 and here re-sequenced some of the same amplicon pools. This allowed for a precise assessment of the improvements. In 2018, sequencing the 511 amplicons of the *Mixed Diptera* sample required one flow cell and we obtained 488 barcodes of which only one lacked ambiguous bases. In 2021, we sequenced on a used flow cell (R10.3) with only ~500 pores, and we obtained 502 barcodes with >98% (496) being free of ambiguous bases. The results obtained for the 2019 amplicon pools were also better. In 2019, one flow cell (R9.4) allowed us to obtain 3,223 barcodes from a pool of amplicons obtained from 4,275 specimens of *Afrotropical Phoridae*. Resequencing weak amplicons increased the total number of barcodes by approximately 500 to 3,762 [33]. Now, one R10.3 flow cell yielded 3,905 barcodes (+143) for the same amplicon pool, while retaining an accuracy of >99.99% and reducing the ambiguities from 0.45% to 0.01%. If progress continues at this pace, we predict that MinION will be the default barcoding tool for most users. This, too, is because all barcoding steps can now be carried out in one laboratory with a modest set of equipment (see Table 4). With MinION being readily available, there is no longer the need to outsource sequencing and/or to wait until enough barcode amplicons have been prepared for an Illumina or PacBio/Sequel flow cell [57]. This eases biodiversity discovery and allows many biologists, government agencies, students, and citizen scientists from around the globe to get involved in biodiversity discovery.

Table 4. Equipment required for MinION barcoding

Required (<500 specimens)	
1	MinION sequencer (preferably Mk1C for basecalling) with Flongle adapter
2	Thermocycler(s)
3	Gel Electrophoresis setup
4	Magnetic Separation Rack

5	Qubit for DNA quantification
6	Standard equipment: Vortex, Mini-centrifuge, pipettes, freezer, fridge
7	Standard laptop or PC
Required (>500 specimens)	
1	Multichannel pipette(s)
Optional but highly desirable	
1	Hula Mixer

This raises the question of how much it costs to sequence a barcode with MinION. There is no straightforward answer because the cost depends on user targets. For example, a user who wants to sequence a pool of 5000 barcodes may want an 80% success rate in order to identify the dominant species in a sample of nuisance midges that occur in mixed-species swarms [58]. Based on Figure 4, only ca. 1.5 million raw MinION reads would be needed. On average, MinION flow cells yield >10 million reads and cost USD 475-900 depending on how many cells are purchased at the same time. Add the library cost of ca. USD 100 and the overall sequencing cost of the project is USD 180-235. This experiment would be expected to yield 4000 barcodes for the 5000 amplicons (4-6 cents/barcode). Given the low cost of 1 million MinION reads ($50-90), we predict that most users will opt for sequencing at a greater depth since this will likely yield several hundred additional barcodes. This will increase the sequencing cost per barcode, because the first 1.5 million reads already recovered barcodes for all strong amplicons. Additional reads will predominantly strengthen read coverage for these amplicons and relatively few reads will be added to the read bins that were too weak to yield barcodes at low coverage; i.e., there are diminishing returns for additional sequencing.

Overall, we thus predict that most users will only multiplex 10,000 amplicons in the same MinION flow cell so that the sequencing cost per specimen would be 0.06-0.10 USD

depending on flow cell cost. Large-scale biodiversity projects can reduce the cost further by switching to sequencing with PromethION, a larger ONT sequencing instrument that can accommodate up to 48 flow cells. PromethION flow cells have 6 times the number of pores for twice the cost so that this switch reduces the sequencing cost by 60% (flow cell capacity: 60,000 barcodes). At the other end of the scale are those users who occasionally need a few hundred barcodes. They can use Flongle flow cells, which are comparatively expensive (0.50 USD) because each flow cell costs $70 and requires a library that is prepared with half the normal reagents (ca. $50).

ONTbarcoder for large-scale species discovery with MinION

We here introduce ONTbarcoder, which runs on a regular laptop using either Windows10, Linux, or Macintosh OS and has a GUI. ONTbarcoder has been extensively tested (>4000 direct comparisons to Sanger and Illumina barcodes) and is designed to yield thousands of barcodes rapidly without impairing accuracy even when encountering low-coverage amplicons. Accuracy is ensured by applying 4 QC criteria related to length and translatability of the barcode. Speed is maintained through the parallelization of most steps on UNIX systems (Mac and Linux; parallelization is restricted to demultiplexing in Windows). ONTbarcoder furthermore allows for updating the parameter file for alignment. This is advisable because MinION continues to evolve quickly. We expect flow cell capacity to increase further and basecalling to improve (see [59]). For example, a new basecaller ("*bonito*") developed by ONT has shown promise by improving raw read accuracy (https://nanoporetech.com/about-us/news/new-research-algorithms-yield-accuracy-gains-nanopore-sequencing). This basecaller is now also available in MinKNOW and our preliminary tests (Flongle: *Mixed Diptera Subsample, Chironomidae*; R10.3: *Palaearctic Phoridae,* 313 bp; bonito version=0.3.6*)* confirm that it yields reads of similar quality as HAC (unpublished data). We expect these regular changes to ONT software to further improve the suitability of ONT sequencers for barcoding.

ONTbarcoder evolved from miniBarcoder, which yielded high quality barcodes based on four different tests covering >8000 barcodes [31, 33, 54, 60], but had two drawbacks that have been fixed in ONTbarcoder. Firstly, we dropped the translation-based error correction that tended to increase the number of Ns. This step used to be essential because indel errors were prevalent in consensus barcodes obtained with older flow cell models. Secondly, ONTbarcoder can be installed by unzipping a file and is easy to maintain on different operating systems. Until now, external dependencies meant that several software packages had to be installed and that only some operating systems were compatible. This has been a major drawback of all MinION bioinformatics pipelines and led Watsa et al. [35] to recommend that bioinformatics training is needed before MinION barcoding could be used in schools (e.g., training in UNIX command-line). With ONTbarcoder, such training will no longer be needed.

MinION has been used for barcoding fungi, animals, and plants and alternative pipelines have been developed [36, 54, 56, 60-66], but there is one fundamental difference between these studies/pipelines and the vision presented here. These studies tended to show that MinION sequencing can be done in the field. Thus only a very small number of specimens were analysed (<150 with the exception of >500 in Chang, Ip et al [60]). The potential use in the field is an attractive feature for time-sensitive samples that could degrade before reaching a lab. However, it is unlikely to help substantially with tackling large-scale biodiversity discovery and monitoring because obtaining few MinION barcodes per flow cell is too expensive. Additionally, the bioinformatic pipelines that were developed for these small-scale projects were not suitable for large-scale, decentralized barcoding. For example, some used ONT's commercial barcoding kit that only allows for multiplexing up to 96 samples in one flow cell [64, 66]; i.e., each amplicon has very high read coverage which influenced the design of the analysis pipelines (e.g. coverage recommendations for ONTrack is 1000x: Maestri, Cosentino et al.[64]). The high coverage requirements also meant that the pipelines were only tested for small numbers of samples (<60: Menegon, Cantaloni et al.

2017, Maestri, Cosetino et al. 2019, Seah, Lim et al. 2020, Sahlin, Lim et al. 2021 [56, 61, 64, 66]) which are unlikely to represent the complexities of large, multiplexed amplicon pools (e.g., nucleotide diversity, uneven coverage).

This concern is confirmed by our test of the most recently introduced bioinformatics pipeline (NGSpeciesID [56]). It requires minibar/qcat and nanofilt, isONclust SPOA, Parasail, and optionally, Medaka [63, 67, 68] and often yields multiple consensus barcodes for the same set of reads because it relies on an intermediate clustering step. To assess the performance of NGSpeciesID, we used the consensus barcode with the highest coverage and only compared the results for those barcodes for which we had reference barcodes obtained with Sanger and Illumina (Table 3). We find that under optimal settings, NGSpeciesID yields 3-23 times the number of erroneous barcodes than ONTbarcoder (Table 3). One reason is that NGSpeciesID does not use ambiguity codes in consensus sequences; i.e., ONTbarcoder will place an “N” when the evidence is ambiguous while NGSpeciesID will opt for one of the nucleotides. In addition, NGSpeciesID does not use barcode length or translatability as QC criteria. This has the downside that the software is more likely to yield erroneous barcodes. For example, NGSpeciesID proposes consensus barcodes for very small read sets for which ONTbarcoder proposes no barcodes because they failed the QC or did not meet the minimum read threshold. An upside of not using barcode length and translatability as QC is that NGSpeciesID can propose consensus barcodes for genes that are non-coding (e.g., ITS) or have high length variability (e.g., ribosomal genes). However, the user should be aware that 3-5% of these barcodes will include errors if the results for *COI* also hold for other genes (see Table 3).

Biodiversity monitoring with MinION barcodes

Despite the widespread use of metabarcoding for analyzing samples consisting of hundreds or thousands of specimens [69, 70], large-scale barcoding of individual specimens remains essential for discovering and describing species. It associate barcodes with individual

voucher specimens, which can be used for further research. This is essential for taxonomic research, which is the only way to fix systematic errors caused by *COI* which lumps recently diverged species and splits species with deep allopatric splits [71]. High quality barcode databases are also important for the analysis of metabarcoding data because they facilitate the identification of numts, heteroplasmy, contaminants and errors. Furthermore, large-scale barcoding will also be critical for developing AI-assisted biodiversity monitoring of invertebrates using images [72]. Such monitoring requires neural networks that are trained with large numbers of images. These images are best obtained from specimens that were identified/grouped into species based on barcodes barcodes. It appears likely that AI-assisted biodiversity monitoring will be the method of choice in the future because it can has the potential to quickly identify and count common species and highlight which specimens may belong to new/rare species [73].

Conclusions

Many biologists would like to have ready access to barcodes without having to send specimens halfway around the world or run large, complex, and expensive molecular laboratories. Many have been impressed by MinION's low cost, portability, and its ability to deliver real-time sequencing, but they were worried about high cost and complicated bioinformatics pipelines. We here demonstrate that these concerns are no longer justified. MinION barcodes obtained by R10.3 flow cells are virtually identical to barcodes obtained with Sanger and Illumina sequencing. Barcoding with MinION is now also cost-effective and the new "ONTbarcoder" software makes it straightforward to analyze the data on a standard laptop. Add the simplified and cheaper methods for obtaining amplicons and biodiversity discovery will become more scalable and accessible to all.

Methods

MinION and Flongle sequencing were here tested for six amplicon pools (Table 5). For two of the pools, *Mixed Diptera* (N=511) and *Afrotropical Phoridae* (N=4,275), we already had

amplicons and comparison barcodes that were obtained with Sanger, Illumina, and older versions of MinION flow cells that used a different chemistry (see below). These two pools were here used to assess the accuracy of barcodes generated using the new MinION flow cell using the R10.3 chemistry. Two additional datasets were used to test the capacity of R10.3 flowcells for mini- and full length barcodes by sequencing barcodes of different lengths for the same specimens and obtained with the same DNA template (*Palaearctic Phoridae,* 658 and 313 bp for ca. 9,930 specimens). Lastly, we tested the performance of Flongle flow cells using a *Chironomidae* dataset (313 bp mini-barcodes for 191 specimens) and a *Mixed Diptera Subsample* (full-length barcodes for 257 specimens) of the aforementioned *Mixed Diptera* amplicon pool for which we had Sanger barcodes for comparison.

The methods for obtaining the reference barcodes with Sanger and Illumina are described in Srivathsan et al. [31, 33]. Briefly, for Sanger sequencing (Mixed Diptera sample set), the same PCR products sequenced with MinION were individually cleaned using SureClean Plus (Bioline, London) and subjected to cycle sequencing using BigDye^{TM}. The products were precipitated using PureSeq (Aline BioSciences, Woburn) and analysed in ABI 3730xl 96 capillary sequences. The resulting chromatograms were edited using Sequencher v4 (GeneCodes, Ann Arbor). For Illumina sequencing of products of the Afrotropical Phoridae dataset, an independent PCR was conducted using the DNA extract for the same specimens to amplify a short 313-bp fragment of COI using 9-bp tagged versions of the primers described by Leray et al. [74]. The products were pooled and sequenced using HiSeq2500 (250 bp PE sequencing). The data processing followed Wang et al.'s protocol. [23] for obtaining a set of consensus barcodes.

Table 5. Datasets used in the study and the corresponding experimental details.

Dataset Name	**Number of specimens**	**Fragment size, primer information**	**Extraction/PCR setup**	**PCR cleanup**	**ONT Library Preparation kit/Flow cell used**
R10.3 Datasets					
Mixed Diptera	511 (257 mixed	658 bp	Extraction Method:	Ampure	SQK-

(see Srivathsan et al. 2018) - Sanger barcodes available	Diptera, 254 Dolichopodidae) 17 negatives	HCO2198, LCO1490 [75]	QuickExtract PCR Mix: Total volume: 20 µl 10× buffer: 2 µl dNTPs (2.5 mM): 1.5 µl Taq polymerase: 0.2 µl BSA (1 mg/ml): 2 µl Primer (5 µM): 2 µl each DNA:2 µl	beads (Beckman Coulter)	LSK110/FLO-MIN111
Afrotropical Phoridae (see Srivathsan et al. 2019) - Illlumina mini-barcodes available	4275 (Phoridae) 45 negatives	658 bp HCO2198, LCO1490 [75]	Extraction Method: QuickExtract PCR Mix: Total volume: 15.16 µl Mastermix (CWBio): 10 µl 25mM MgCl2: 0.16 µl BSA (1mg/ml): 2 µl Primer (10µM): 1 µl each DNA: 1 µl	Sera-Mag beads (GE Healthcare Life Sciences) in PEG	SQK-LSK109/FLO-MIN111
Palaearctic Phoridae (658)	9,929 (Phoridae) 105 negatives	658 bp jgHCO2198, LCO1490 [75, 76]	Extraction Method: HotSHOT PCR Mix: Total volume: 16 µl Mastermix (CWBio): 7 µl BSA (1mg/ml): 1 µl Primer (10µM): 1 µl each DNA: 6 µl	Ampure beads (Beckman Coulter)	SQK-LSK110/FLO-MIN111
Palaearctic Phoridae (313)	9,932 (Phoridae) 106 negatives	313 bp m1COlintF, jgHCO2198 [74, 76]	Extraction Method: HotSHOT PCR Mix: Total volume: 14 µl Mastermix (CWBio): 7 µl BSA (1mg/ml): 1 µl Primer (10µM): 1 µl each DNA: 4 µl	Ampure beads (Beckman Coulter)	SQK-LSK110/FLO-MIN111
Flongle Datasets					
Mixed Diptera subsample (see Srivathsan et al. 2018) - Sanger barcodes available	257 7 negatives	See “Mixed Diptera” entry for R10.3	See “Mixed Diptera” entry for R10.3	Ampure beads (Beckman Coulter)	SQK-LSK109/Flongle
Chironomidae	191 (Chironomidae) 1 negative	313 bp m1COlintF, jgHCO2198 [74,76]	Extraction Method: HotSHOT PCR Mix: Total volume: 14 µl Mastermix (CWBio): 7 µl BSA (1mg/ml): 1 µl Primer (10µM): 1 µl each DNA: 4 µl	Ampure beads (Beckman Coulter)	SQK-LSK109/Flongle

1: DNA extraction

For all newly barcoded specimens, we used DNA template obtained with 10-15 µL HotSHOT per specimen [77], but other buffers like PBS could have also been used (see [78]). Small specimens were submerged within the well of a microplate while larger specimens were placed head-first into the well. Note that the specimen need not be entirely submerged in HotSHOT. The placement of 95 specimens takes approximately 17 minutes and the DNA is

obtained within 20 minutes in a thermocycler via two heating steps [77] (https://www.youtube.com/watch?v=y1qGzL5PraQ&t=3s). After neutralization, >20 µl of template is available for amplifying *COI* and the voucher can be recovered. The advantages of HotSHOT is low cost and speed. The disadvantages are fast degeneration of the leftover template within days. Some alternatives to DNA extraction with HotSHOT are described in Table 6. We consider them too costly or time-consuming given that *COI* is a mitochondrial gene and thus naturally enriched. Indeed, the small mitochondrial genome (16 kbp) usually contributes 0.5-5% of the DNA in a genomic extraction [79, 80]. Therefore obtaining sufficient template for DNA barcoding should take <20 minutes, does not require DNA purification, and should cost essentially nothing as long as the specimens contain DNA of reasonable quality (e.g., <20 year old Malaise trap samples).

Table 6. Alternative DNA extraction methods

Commonly used alternative DNA extraction methods	**advantages**	**disadvantages**
"directPCR": "contaminating" a PCR reaction with the DNA of the target organism by adding the entire specimen or a tissue sample into the PCR reagent mix (Wong, Tay et al. 2014).	• No cost • No waiting time obtaining for template	• time-consuming when subsampling is needed (antenna, leg) • low success rate for heavily sclerotized specimens • no DNA template left after PCR
Commercial DNA extraction buffers: e.g., QuickExtract: 10 µl sufficient for obtaining DNA template from most insect specimens (Srivathsan, Hartop et al. 2019)	• long shelf life of buffers • template stays viable for weeks • additional DNA can be obtained through re-extraction of specimen	• moderate costs (<0.20 USD) • DNA in leftover template degrades within weeks/months
Commercial DNA extraction kits: e.g., DNeasy Blood & Tissue Kits:	• template is stable	• high cost (>1 USD) • time-consuming

2: Amplification of COI via PCR

Obtaining amplicons for DNA barcodes does not require high-fidelity polymerases, which are mostly needed for amplifying low copy-number nuclear genes based on low-concentration template. Standard polymerases are sufficient. We exclusively used CWBio 2x master mix in 14-16 ul PCR reactions (see Table 5 for details). The set-up of one plate requires <15 minutes (https://www.youtube.com/watch?v=NxYOvZGhD0E&t=5s). In order to save time, only a small number of reactions per microplate need to be checked via gel electrophoresis (N=8-12, including the negative control). Running out additional amplicons is not necessary because failed amplicons do not add to the MinION sequencing cost and can be recovered via re-sequencing [33] or re-amplification.

Amplicon sequencing with all second and third generation sequencing technologies (including MinION) involves amplicon pools. This means that the amplicon for each specimen has to be tagged/indexed/barcoded with short DNA sequences at the 5' ends of the amplicons. This allows for the assignment of each read obtained during sequencing to a specific specimen during the "demultiplexing" bioinformatics step. We use 13 bp tags that are distinct (>4bp from each other including insertions/deletions) and lack homopolymers (see supplementary materials). This tag length is a compromise given that longer tags have the disadvantage of reducing PCR success rates [33] while having the advantage of increasing the proportion of reads that can be demultiplexed with confidence.

Numerous dual-PCR tagging techniques for amplicons have been described in the literature [29, 64, 81, 82], but we only use single-PCR tagging [28]. It is here described for a microplate with 96 templates, but the same principle can be applied to strip tubes or partial microplates. What is needed is a 96-well primer plate where each well contains a differently tagged reverse primer. This "primer plate" can yield 96 unique combinations of primers once the 96 reverse primers are combined with one forward primer (f-primer x 96 differently tagged r-primers = 96 unique combinations). This also means that if one purchases 105

differently tagged forward primers, one can individually tag 10,800 specimens (105 x 96= 10,800 amplicons). This is the number of amplicons that we consider appropriate for a MinION flow cell (R10.3; see Results). In the laboratory we assign the tag combinations as follows. For each plate with 96 PCR reactions, we add the same f-primer to a tube with the PCR master mix to be used for the entire plate. We then dispense the “f-primed” master mix into the 96-wells. Afterwards, we use a multichannel pipette to add the DNA template and the tagged r-primers from the r-primer plate into the PCR plate. All 96 samples in the plate now have a unique combination of tagged primers because they only share the same tagged forward primer. This makes tracking of tag combinations simple because each PCR plate has its own tagged f-primer, while the r-primer is consistently tied to well position. Each plate has a negative control that is used to detect contamination. The tagging information for each plate is recorded in the demultiplexing file that is later used to demultiplex the reads obtained during sequencing (see supplement for an example).

We prefer single-PCR tagging over two-PCR tagging [29, 64, 81, 82] because it is cheaper (requires half the amount of primer), less error-prone (fewer PCR reactions and cycles), and saves time (no need to clean-up the first-round amplicons). The only downsides of single-PCR tagging are an initially higher investment in primers and the need to manage the primer stock more carefully because it is used for a longer time. Long-term storage should thus be at -80°C and the number of freeze-thaw cycles should be kept low (<10).

3. Amplicon sequencing

PCR is followed by pooling, purification of the amplicons via the removal of unused PCR reagents, the adjustment of DNA concentration, and sequencing. We only pool 1 µl of each PCR product. The pools were cleaned using SPRI bead-based clean-up with Ampure (Beckman Coulter) beads, but Kapa beads (Roche) or the more cost-effective Sera-Mag beads (GE Healthcare Life Sciences) in PEG [83] are also viable options [33]. For barcodes longer than 300 bp, we recommend the use of a 0.5X ratio for Ampure beads since it

removes a larger proportion of primers and primer dimers. However, this ratio is only suitable if yield is not a concern (e.g., pools consisting of many and/or high concentration amplicons). Increasing the ratio to 0.7-1X will improve yield but render the clean-up less effective. The pooling and clean-up of three 96-well plates takes about 40 minutes (https://www.youtube.com/watch?v=YKVWEvcSw6A), but the time per plate is lower when large numbers of amplicons are pooled. Amplicon pools containing large numbers of amplicons may require multiple rounds of clean-up, but only a small subset of the initial pool has to be purified because most library preparation kits require only small amounts of DNA. We confirm the success of the clean-up procedures via gel electrophoresis, which should show only one strong band of expected length. After the clean-up, the pooled DNA concentration is measured in order to use an appropriate amount of DNA for library preparation. We use Qubit, but less precise techniques are probably also suitable.

Obtaining a cleaned amplicon pool according to the outlined protocol is not time consuming, but many studies retain “old Sanger sequencing habits”. For example, they use gel electrophoresis for each PCR reaction to test whether an amplicon has been obtained. Afterwards, they clean and measure all amplicons – one at a time – for normalization (often with very expensive techniques: Ampure beads: [64]; TapeStation, BioAnalyzer, Qubit:[66]). This is presumably done to obtain a pool of amplicons where each has equal representation. However, reads are cheap while the clean-up and measurement of many amplicons is expensive and unnecessary because weak products that failed to yield a barcode in the first sequencing run can be re-sequenced [33].

4. Library preparation and sequencing

We prepare our MinION libraries using ligation-based kits and 200 ng of DNA for full flow cells and 100 ng for Flongle (see Table 6 for details). We generally follow kit instructions, but exclude the FFPE DNA repair mix in the end-repair reaction, as this is mostly needed for formalin-fixed, paraffin-embedded samples. The reaction volumes for the R10.3 flow cell

libraries consist of 45 µl of DNA, 7 µl of Ultra II End-prep reaction buffer (New England Biolabs), 3 µl of Ultra II End Prep enzyme mix (New England Biolabs) and 5 µl of molecular grade water. For the Flongle, only half of the reagents are used to obtain a total volume of 30 µl. We further modify the Ampure ratio to 1x for all steps as DNA barcodes are short whereas the recommended ratio in the manual is for longer DNA fragments. The libraries for the experiments in this study were loaded and sequenced on a MinION Mk 1B. Data capture involved a MinIT or a Macintosh computer that meets the IT specifications recommended by ONT. The bases were called using Guppy (versions provided in Table 1), under the high-accuracy model in MinIT taking advantage of its GPU.

We here only used MinION and Flongle flow cells for barcoding. This preference was based on four considerations: (1) Scaling; i.e., ability to accommodate projects of different sizes, (2) turnaround times, (3) cost, and (4) amplicon length. Flongle can be used for small amplicon pools (<300 products) because it has low fixed costs per experiment (library and flow cell: ca. $120 USD) and the turnaround time is fast, so the MinION Flongle is arguably the best sequencing option for small barcoding projects with > 50 barcodes. Full MinION flow cells also have fast turnaround times, but the minimum run cost is closer to 1000 USD for most users (flow cell cost drops with bulk purchase), so this option only becomes more cost-effective than Flongle when >1800 amplicons are sequenced. As illustrated in the Results section of the manuscript, one MinION flow cell can comfortably sequence 10,000 amplicons.

Alternatives to MinION/Flongle are sequencing barcode amplicons with Sanger, Illumina (Wang, Srivathsan et al. 2018), or PacBio (e.g., Sequel: Hebert, Braukmann et al. [29]). Sanger sequencing has fast turnaround times but high sequencing cost per amplicon ($3-4 USD). PacBio's Sequel flow cells have a similar capacities as full MinION flow cells [29] and the consumable costs are also similar for most users who have to outsource Sequel sequencing due to the high instrument cost for PacBio. However, Sequel does not allow for

flexible scaling like Flongle/MinION and most users will have to wait several weeks until the data are returned from the service provider. By far the most cost-effective sequencing method for barcodes is Illumina's NovaSeq sequencing. However, the fixed costs for library and lanes are high (3000-4000 USD), but each flow cell yields 800 million reads which can comfortably sequence 800,000 barcodes at a cost of < $0.01 USD per barcode. Note that Illumina reads are only suitable for generating mini-barcodes of up to 420 bp length (using 250bp PE sequencing using SP flow cell). "Full-length" *COI* barcode (658 bp) can only be obtained by sequencing two amplicons per specimen.

5. Bioinformatics: Development and application of ONTbarcoder

One of the most significant barriers to widespread barcoding with MinION has been complex bioinformatics pipelines that were needed for fixing the high error rates of ONT reads. However, after obtaining data from a new R10.3 flow cell, we noticed major improvements in read quality, the total number of raw reads, and the number of demultiplexed reads. This led to the development of "ONTbarcoder", which has a graphical user interface (GUI) and is suitable for all major operating systems (Linux, Mac OS, Windows10). The use of the software is illustrated in a video tutorial: https://www.youtube.com/channel/UC1WowokomhQJRc71FmsUAcg.

ONTbarcoder. ONTbarcoder (available at: https://github.com/asrivathsan/ONTbarcoder) has three modules. (a) The first is a demultiplexing module which assigns reads to specimen-specific bins. (b) The second is a barcode calling module which reconstructs the barcodes based on the reads in each specimen bin. (c) The third is a barcode comparison module that allows for comparing barcodes obtained via different software and software settings.

a. Demultiplexing. In order to obtain barcodes three pieces of information and two files have to be provided to ONTbarcoder via the GUI: (1) primer sequence, (2) expected fragment length, and (3) demultiplexing information (=tag combination for each specimen). The latter

is summarized in a demultiplexing file (see supplementary information for format). The only other required file is the FASTQ file obtained from MinKNOW/Guppy after basecalling. Demultiplexing by ONTbarcoder starts by analyzing the read length distribution in the FASTQ file. Only those reads that meet the read length threshold are demultiplexed (default= 658 bp corresponding to metazoan COI barcode). Technically, the threshold should be the amplicon length plus the length of both tagged primers, but ONT reads have indel errors such that they are occasionally too short or too long. We therefore advise to specify the amplicon length without primer&tag as threshold. ONTbarcoder will split reads that are twice the expected fragment length into two parts whose lengths are determined based fragment size, primer and tag lengths, and a window to account for indel errors (default=100 bp).

Once all reads with a suitable length for demultiplexing have been identified, ONTbarcoder finds the primers via sequence alignment of the primer sequence to the reads (using python library *edlib*). Up to 10 deviations from the primer sequence are allowed because this step is only needed for determining the primer location and orientation within the read. For demultiplexing based on the tags, the flanking region of the primer sequence is retrieved whereby the number of retrieved bases is equal to the user-specified tag length. The flanking sequences are then matched against the tags from the user-provided tag combinations (demultiplexing file). In order to account for sequencing errors, not only exact matches are accepted, but also matches that differ by up to 2 bps from the tag sequence (substitutions/insertions/deletions). Note that accepting tag variants does not lead to demultiplexing error because all tags differ by >4 bp. All reads with the same tag combination thus identified belong to the same specimen and are pooled into the same bin. To increase efficiency, demultiplexing is parallelized and the search space for primers and tags are restricted to the beginning and end of the reads (window is user-specified).

b. Barcode calling: Barcode calling uses the reads within each specimen-specific bin to reconstruct the barcode sequence. The reads are aligned to each other and a consensus sequence is called. Barcode calling is done in three phases: “Consensus by Length”, “Consensus by Similarity” and “Consensus by barcode comparison”. The user can opt to only use some of these methods.

“Consensus by Length” is the main barcode calling mode and relies on efficient alignment in order to provide reasonable speed for thousands bins each containing many reads. ONTbarcoder delivers speed by using an iterative approach that gradually increases the number of reads (“coverage”) per bin that is used during alignment. However, reconstructing barcodes based on few reads could lead to errors which are weeded out by ONTbarcoder by applying four Quality Control (QC) criteria. The first three QC criteria are applied immediately after the consensus sequence has been called: (1) the barcode must be translatable, (2) it has to match the user-specified barcode length, and (3) the barcode has to be free of ambiguous bases (“N”). To increase the chance of finding a barcode that meets all three criteria, we subsample the reads in each bin by read length (thus the name “Consensus by Length”); i.e., initially only those reads closest to the expected length of the barcode are used. For example, if the user specified coverage=25x for a 658 bp barcode, ONTbarcoder would only use the 25 reads that have the closest match to 658 bp. The fourth QC measure is only applied to barcodes that have already met the first three QC criteria. A multiple sequence alignment (MSA) is built for the barcodes obtained from the amplicon pool, and any barcode that causes the insertion of gaps in the MSA is rejected. Note that if the user suspects that barcodes of different length are in the amplicon pool, the initial analysis should use the dominant barcode length. The remaining barcodes can then be recovered by re-analyzing all data or only the failed read bins (“remaining”, see below) and bins that yielded barcodes that had to be “fixed”. These bins can be reanalyzed using a different pre-set barcode length.

“Consensus by Similarity”. Barcodes that failed the QC during the “Consensus by Length” stage are often very close to the expected length and have few ambiguous bases, and/or cause few gaps in the MSA. These “preliminary barcodes” can be improved through “Consensus by Similarity”. This method eliminates outlier reads from the read pool in the bins. Such reads can differ considerably (see below) from the signal of the consensus barcode and ONTbarcoder identifies them by sorting all reads by similarity to the preliminary barcode. Only the top 100 reads (this default can be changed) that differ by <10% from the preliminary barcode are retained and used for calling the barcodes again using the same techniques described “Consensus by Length” (including the same QC criteria). This improvement step converts many preliminary barcodes found during “Consensus by Length” into barcodes that pass all four QC criteria by filling/removing indels or resolving an ambiguous base.

“Consensus by barcode comparison”. The remaining preliminary barcodes that still failed to convert into QC-compliant barcodes tend to be based on read bins with low coverage, but some can yield good barcodes after subjecting them to a further improvement step that fixes the remaining errors. ONTbarcoder identifies these errors by finding the 20 most similar QC-compliant barcodes that have already been reconstructed for the other amplicons. The 21 sequences are aligned and ONTbarcoder finds the errors because they cause insertions and deletions in the MSA. Insertions are deleted, gaps are filled with ambiguous bases (“N”), but mismatches are retained. The number and kinds of “fixes” are recorded and added to the FASTA header of the barcode. Large number of fixes imply that the barcode should not be used (see below). “Rare taxa are disadvantaged by this method, but the barcode for very few if any will ever reach the “Consensus by barcode comparison” stage because most/all will be resolved earlier. We added “consensus by barcode comparison” because it helps with resolving weak barcodes for specimens that represent abundant species.

Output. ONTbarcoder produces a summary table (Outputtable.csv) and FASTA files that contain the different classes of barcodes. Each barcode header contains information on coverage used for barcode calling, coverage of the specimen bin, length of the barcode, number of ambiguities and number of indels fixed. Five sets of barcodes are provided, here discussed in the order of barcode quality: (1) “QC_compliant”: The barcodes in this set satisfy all four QC criteria without correction and are the highest quality barcodes. (2) “Filtered_barcodes”: this file contains the barcodes that are translatable, have <1% ambiguities and have up to 5 indels fixed during the last step of the bioinformatics pipeline. These filtering thresholds were calibrated based on two datasets for which we have Sanger/Illumina barcodes and the resulting barcodes are found to be highly accurate. Note that the file with filtered barcodes also includes the QC_compliant barcodes and that all results discussed in this manuscript are based on filtered barcodes given that they are of much higher quality than the average barcode in BOLDSystems (assessment in Srivathsan, Baloğlu et al [31]).

The remaining files include barcodes of lesser and/or suspect quality. (3) “Fixed_barcodes_XtoY”: these files contain barcodes that had indel errors fixed and are grouped by the number of errors fixed. Only the barcodes with 1-5 errors overlap with Filtered barcodes file, if they have <1% ambiguities. (4) “Allbarcodes”: this file contains all barcodes in sets (1)-(3). (5) “Remaining”: these are barcodes that fail to either translate or are not of predicted length. Note that all barcodes should be checked via BLAST against comprehensive databases in order to detect lab contamination. There are several online tools available for this and we recommend the use of GBIF sequence ID tool (https://www.gbif.org/tools/sequence-id) which gives straightforward output including a taxonomic summary.

The output folder also includes the FASTA files that were used for alignment and barcode calling. The raw read bins are in the “demultiplexed” folder, while the resampled bins (by

length, coverage, and similarity) are in their respective subfolders named after the search step. Note that the raw reads are encoded to contain information on the orientation of the sequence and thus cannot be directly used in other software without modifications (see ONTbarcoder manual on Github). Lastly, for each barcode FASTA file (1-5), there are folders with the files that were used to call the barcodes. This means that the user can, for example, reanalyze those bins that yielded barcodes with high numbers of ambiguous bases. Lastly a “runsummary.xlsx” document allows the user to explore the details of the barcodes obtained at every step of the pipeline.

Algorithms. ONTbarcoder uses the following published algorithms. All alignments utilize MAFFTv7 (Katoh and Standley 2013). The MinION reads are aligned using an approach similar to lamassemble [84] with parameters optimized for nanopore data by “last-train” [85] which accounts for strand specific error biases. The MAFFT parameters can be modified in the “parfile” supplied with the software which helps with adjusting the values given the rapidly changing nanopore technology. All remaining MSAs in the pipeline (e.g., of preliminary barcodes) use MAFFT’s default settings. All read and sequence similarities are determined with the *edlib* python library under the Needle-Wunsch (“NW”) setting, while primer search is using the infix options (“HW”). All consensus sequences are called from within the software. This is initially done based on a minimum frequency of 0.3 for each position. This threshold was empirically determined based on datasets where MinION barcodes can be compared to Sanger/Illumina barcodes. The threshold is applied as follows. All sites where >70% of the reads have a gap are deleted. For the remaining sites, ONTbarcoder accepts those consensus bases that are found in at least >30% of the reads. If no base/multiple bases reach this threshold, an “N” is inserted. To avoid reliance on a single threshold, ONTbarcoder allows the user to change the consensus calling threshold from 0.2 to 0.5 for all barcodes that fail the QC criteria at 0.3 frequency. However, barcodes called at different frequencies are only accepted if they pass the first three QC criteria and

are identical. If no such barcode is found, the 0.3 frequency consensus barcode is used for further processing.

c. Barcode comparison. Many users may want to call their barcodes under different settings and then compare barcode sets. The ONTbarcoder GUI simplifies such comparisons. A set of barcodes can be dragged into the window and the user can select a barcode set as the reference. The barcode comparisons are conducted using *edlib* library. The barcodes in the sets are compared and classified into three categories: “identical” where sequences are a perfect match and lack ambiguities, “compatible” where the sequences only differ by ambiguities, and “incorrect” where the sequences differ by at least one base pair. Several output files are provided. A summary sheet, a FASTA file each for “identical”, “compatible”, and the sequences only found in one dataset. Lastly, there is a folder with FASTA files containing the different barcodes for each incompatible set of sequences. This module can be used for either comparing set(s) of barcodes to reference sequences, or for comparing barcode sets against each other. It furthermore allows for pairwise comparisons and comparisons of multiple sets in an all-vs-all manner. This module was here used to get the final accuracy values presented in Table 2.

Quality of Flongle and MinION barcodes

We first used ONTbarcoder to analyze the data for all six datasets by analyzing all specimen-specific read bins at different coverages (5-200x in steps of 5x). This means that the barcodes for a bin with 27 reads were called five times at 5x, 10x, 15x, 20x, and 25x coverages while bins with >200x were analyzed 40 times at 5x increments. Instead of using conventional rarefaction via random subsampling reads, we used the first reads provided by the flow cell because this accurately reflects how the data accumulated during the sequencing run and how many barcodes would have been obtained if the run had been stopped early. This rarefaction approach also allowed for mapping the barcode success rates against either coverage or time.

In order to obtain a “best” estimate for how many barcodes can be obtained, we also carried out one analysis at 200x coverage with the maximum number of “Comparison by Similarity” reads set to 100. All analyses produced a “filtered” set of barcodes (barcodes with <1% Ns and up to 5 fixes) that were used for assessing the accuracy and quality via comparison with Sanger and Illumina barcodes for *Mixed Diptera (MinION R10.3)*, *Afrotropical Phoridae (MinION R10.3)*, and *Mixed Diptera Subsample (Flongle R9.4).* For the comparisons of the barcode sets obtained at the various coverages, we used MAFFT and the assess_corrected_barcode.py script in miniBarcoder (Srivathsan et al. 2019).

6. Bioinformatics: Application of minibar and NGSpeciesID

We compared the barcodes obtained with ONTbarcoder with those obtained with the recently published NGSpeciesID [56]. This comparison was carried out for the datasets that have the reference barcodes obtained with Sanger and Illumina (*Mixed Diptera (MinION R10.3)*, *Afrotropical Phoridae (MinION R10.3)*, and *Mixed Diptera Subsample (Flongle R9.4)*). Two comparisons were made. Firstly, we used the same demultiplexed reads that were used for calling the barcodes using ONTbarcoder. Secondly, we demultiplexed the data using *minibar* (git commit: 938ae51) and applied NGSpeciesID (Git commit: 24afc6c) for consensus barcode calling. This approach allowed for software comparisons for consensus calling and demultiplexing. *minibar* was run using the parameters (-e 2 -E 10 -T), i.e. maximum number of errors in tag region was set to 2, in primer alignment was set to 10 and primer and tags were trimmed from the sequences. NGSpeciesID was run with both full datasets as well as by subsampling the data to 200 reads (--sample_size 200) to keep it comparable with our analysis using ONTbarcoder. Other parameters settings were: intended target length = 658 (--m) and maximum deviation from target length = 50 (--s). Lastly, for the R10.3 datasets, the medaka model was specified by using –model r103_min_high_g345, which was the only available for R10.3.

Acknowledgements

We would like to thank John T. Longino and Michael Branstetter for providing valuable comments on the manuscript. For the Palaearctic phorid samples, we would like to thank Dave Karlsson, the Swedish Insect Inventory Project, and the crew at Station Linné that sorted out the phorids. We would also like to thank Wan Ting Lee for help with molecular work, and the numerous staff, students and interns who have contributed to the establishment of the pipeline in the NUS laboratory. We would also like to acknowledge Suphavilai Chayaporn and Niranjan Nagarajan from Genome Institute of Singapore for their help with basecalling using bonito. This work was supported by a Ministry of Education grant on biodiversity discovery (R-154-000-A22-112).

Declarations: Software and test dataset availability

ONTbarcoder is available at https://github.com/asrivathsan/ONTbarcoder, which also contains the link to download the raw data and demultiplexing files. The manual for the software is included in the repository https://github.com/asrivathsan/ONTbarcoder/blob/main/ONTBarcoder_manual.pdf. The videos tutorials can be found in the YouTube channel Integrative Biodiversity Discovery: https://www.youtube.com/channel/UC1WowokomhQJRc71FmsUAcg. The datasets have been uploaded to NCBI, under BioProject: PRJNA745481, SRA accession numbers: SRR15185964, SRR15098600, SRR15098599, SRR15188571, SRR15188570, SRR15188569. The datasets, demultiplexing files and reference barcodes are also available via doi:10.5281/zenodo.5115258.

Literature cited

1. Hebert PDN, Cywinska SL, Ball SL, DeWaard JR: **Biological identifications through DNA Barcodes**. *Proc Biol Sci* 2003, **270**(1512):313-321.
2. Tautz D, Arctander P, Minelli A, Thomas RH, Vogler AP: **A plea for DNA taxonomy**. *Trends in Ecology & Evolution* 2003, **18**(2):70-74.

885 3. Meier R: **DNA sequences in taxonomy - Opportunities and challenges**. In: *The New*
886 *Taxonomy Systematics Association Special Volume*. Edited by Wheeler QD. New York: CRC
887 Press; 2008: 95-128.
888 4. Ivanova NV, DeWaard JR, Hebert PD: **An inexpensive, automation-friendly protocol for**
889 **recovering high-quality DNA**. *Molecular Ecology Notes* 2006, **6**(4):998-1002.
890 5. Ivanova NV, Borisenko AV, Hebert PD: **Express barcodes: racing from specimen to**
891 **identification**. *Molecular ecology resources* 2009, **9 Suppl s1**:35-41.
892 6. Meier R, Blaimer B, Buenaventura E, Hartop E, Von Rintelen T, Srivathsan A, Yeo D: **A re-**
893 **analysis of the data in Sharkey et al.'s (2021) minimalist revision reveals that BINs do not**
894 **deserve names, but BOLD Systems needs a stronger commitment to open science**. *BioRxiv*
895 2021, **doi**:10.1101/2021.1104.1128.441626.
896 7. Stork NE, McBroom J, Gely C, Hamilton J: **New approaches narrow global species estimates**
897 **for beetles, insects, and terrestrial arthropods**. *PNAS* 2015, **112**(24):7519-7523.
898 8. Yeo D, Srivathsan A, Meier R: **Longer is Not Always Better: Optimizing Barcode Length for**
899 **Large-Scale Species Discovery and Identification**. *Systematic Biology* 2020, **69**(5):999-1015.
900 9. **World Economic Forum. The Global Risks Report 2020.**
901 [https://www.weforum.org/reports/the-global-risks-report-2020]
902 10. Re S: **Biodiversity and Ecosystem Services A business case for re/insurance**. In. Zurich: Swiss
903 Re Management Ltd.; 2020.
904 11. Abrego NT, Roslin T, Huotari Y, Ji NM, Schmidt NM, Wang J, Yu D, Ovaskainen O: **Accounting**
905 **for species interactions is necessary for predicting how arctic arthropod communities**
906 **respond to climate change**. *Ecography* 2021, **doi**:10.1111/ecog.05547.
907 12. Kwong S, Srivathsan A, Meier R: **An update on DNA barcoding: low species coverage and**
908 **numerous unidentified sequences**. *Cladistics* 2012, **28**(6):639-644.
909 13. **Dark taxa: GenBank in a post-taxonomic world** [https://iphylo.blogspot.com/2011/04/dark-
910 taxa-genbank-in-post-taxonomic.html]
911 14. Barrett RDH, Hebert PD: **Identifying spiders through DNA barcodes**. *Canadian Journal of*
912 *Zoology* 2005, **83**:481-491.
913 15. Hendrich L, Pons J, Ribera I, Balke M: **Mitochondrial Cox1 sequence data reliably uncover**
914 **patterns of insect diversity but suffer from high lineage-idiosyncratic error rates**. *PLoS One*
915 2010, **5**(12):e14448.
916 16. Hebert PD, DeWaard JR, Zakharov EV, Prosser SWJ, Sones JE, McKeown JTA, Mantle B, La
917 Salle J: **A DNA 'Barcode Blitz': Rapid Digitization and Sequencing of a Natural History**
918 **Collection**. *PLoS One* 2013, **8**(7):e68535.
919 17. Ng'endo RN, Osiemo ZB, Brandl R: **DNA Barcodes for Species Identification in the**
920 **Hyperdiverse Ant Genus Pheidole (Formicidae: Myrmicinae)**. *Journal of Insect Science* 2013,
921 **13**:27.
922 18. Hebert PD, Ratnasingham S, Zakharov EV, Tefler AC, Levesque-Beaudin M, Milton A,
923 Pedersen S, Jannetta P, DeWaard JR: **Counting animal species with DNA barcodes: Canadian**
924 **insects**. *Philosophical Transactions of the Royal Society B: Biological Sciences* 2016,
925 **371**:20150333.
926 19. Thormann B, Ahrens D, Armijos DM, Peters MK, Wagner T: **Exploring the leaf beetle fauna**
927 **(Coleoptera: Chrysomelidae) of an Ecuadorian mountain forest using DNA barcoding**. *PLoS*
928 *One* 2016, **11**(2):e0148268.
929 20. Knox MA, Hogg ID, Pilditch CA, Garcia-R JC, Hebert PDN, Steinke D: **Contrasting patterns of**
930 **genetic differentiation for deep-sea amphipod taxa along New Zealand's continental**
931 **margins**. *Deep Sea Research Part I: Oceanographic Research Papers* 2020, **162**:103323.
932 21. Krell FT: **Parataxonomy vs. taxonomy in biodiversity studies – pitfalls and applicability of**
933 **'morphospecies' sorting.** . *Biodiversity and Conservation* 2004, **13**(4):795-812.

934 22. Stribling JB, Pavlik KL, Holdsworth SM, Leppo EW: **Data quality, performance, and**
935 **uncertainty in taxonomic identification for biological assessments**. *Journal of the North*
936 *American Benthological Society* 2008, **27**(4):906-919.
937 23. Wang WY, Srivathsan A, Foo M, Yamane SK, Meier R: **Sorting specimen-rich invertebrate**
938 **samples with cost-effective NGS barcodes: Validating a reverse workflow for specimen**
939 **processing**. *Molecular ecology resources* 2018, **18**(3):490-501.
940 24. Puillandre N, Modica MV, Zhang Y, Sirovich L, Boisselier MC, Cuaud C, Holford M, Samadi S:
941 **Large-scale species delimitation method for hyperdiverse groups.** *Molecular Ecology* 2012,
942 **21**:2671-2691.
943 25. Hartop E, Srivathsan A, Ronquist F, Meier R: **Large-scale Integrative Taxonomy (LIT):**
944 **resolving the data conundrum for dark taxa.** . *BioRxiv* 2021,
945 **doi**:10.1101/2021.1104.1113.439467.
946 26. Shokralla S, Gibson JF, Nikbakht H, Janzen D, Hallwachs W, Hajibabaei M: **Next-generation**
947 **DNA barcoding: using next-generation sequencing to enhance and accelerate DNA barcode**
948 **capture from single specimens**. *Molecular ecology resources* 2014, **14**(5):892-901.
949 27. Shokralla S, Porter TM, Gibson JF, Dobosz R, Janzen D, Hallwachs W, Golding GB, Hajibabaei
950 M: **Massively parallel multiplex DNA sequencing for specimen identification using an**
951 **Illumina MiSeq platform**. *Scientific Reports* 2015, **5**:9687.
952 28. Meier R, Wong W, Srivathsan A, Foo M: **$1 DNA barcodes for reconstructing complex**
953 **phenomes and finding rare species in specimen-rich samples**. *Cladistics* 2016, **32**:100-110.
954 29. Hebert PD, Braukmann TWA, Prosser SWJ, Ratnasingham S, deWaard JR, Ivanova NV, Janzen
955 D, Hallwachs W, Naik S, Sones JE *et al*: **A Sequel to Sanger: amplicon sequencing that scales**.
956 *BMC Genomics* 2018, **19**:219.
957 30. Krehenwinkel H, Kennedy SR, Rueda A, Lam A, Gillespie RG: **Scaling up DNA barcoding –**
958 **Primer sets for simple and cost efficient arthropod systematics by multiplex PCR and**
959 **Illumina amplicon sequencing**. *Methods in Ecology and Evolution* 2018, **9**(11):2181-2193.
960 31. Srivathsan A, Baloglu B, Wang W, Tan WX, Bertrand D, Ng AHQ, Boey EJH, Koh JJY, Nagarajan
961 N, Meier R: **A MinION-based pipeline for fast and cost-effective DNA barcoding**. *Molecular*
962 *ecology resources* 2018, **18**(5):1035-1049.
963 32. Yeo D, Srivathsan A, Puniamoorthy J, Foo M, Grootaert P, Chan L, Guenard B, Damken C,
964 Wahab RA, Ang YC *et al*: **Mangroves are an overlooked hotspot of insect diversity despite**
965 **low plant diversity**. *BMC Biology* 2021, **doi**:10.1186/s12915-12021-01088-z.
966 33. Srivathsan A, Hartop E, Puniamoorthy J, Lee WT, Kutty SN, Kurina O, Meier R: **Rapid, large-**
967 **scale species discovery in hyperdiverse taxa using 1D MinION sequencing.** *BMC Biology*
968 2019, **17**(1):96.
969 34. Ponder W, Lunney D: **The Other 99% - the Conservation and Biodiversity of Invertebrates**.
970 Sydney: Transactions of the Royal Zoological Society of New South Wales; 1999.
971 35. Watsa M, Erkenswick GA, Pomerantz a, Prost S: **Portable sequencing as a teaching tool in**
972 **conservation and biodiversity research**. *PLoS Biology* 2020, **18**(4):e3000667.
973 36. Pomerantz A, Peñafel A, Arteaga A, Bustamante L, Pichardo F, Coloma LA, Barrio-Amorós CL,
974 Salazar-Valenzuela D, Prost S: **Real-time DNA barcoding in a rainforest using nanopore**
975 **sequencing: opportunities for rapid biodiversity assessments and local capacity building**.
976 *2018* 2018, **7**(4):giy033.
977 37. Marshall SA: **Field photography and the democratization of arthropod taxonomy** *American*
978 *Entomologist* 2008, **54**(4):207-210.
979 38. Dunn RR, Beasley DE: **Democratizing evolutionary biology, lessons from insects**. *Current*
980 *Opinion in Insect Science* 2016, **18**:89-92.
981 39. Baloglu B, Clews E, Meier R: **NGS barcoding reveals high resistance of a hyperdiverse**
982 **chironomid (Diptera) swamp fauna against invasion from adjacent freshwater reservoirs**.
983 *Front Zool* 2018, **15**:31.

40. 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.
41. Lim NK, Tay YC, Srivathsan A, Tan JW, Kwik JT, Baloglu B, Meier R, Yeo DC: **Next-generation freshwater bioassessment: eDNA metabarcoding with a conserved metazoan primer reveals species-rich and reservoir-specific communities**. *Royal Society open science* 2016, **3**(11):160635.
42. Srivathsan A, Nagarajan N, Meier R: **Boosting natural history research via metagenomic clean-up of crowdsourced feces**. *PLoS Biology* 2019, **17**(11):e3000517.
43. Grootaert P: **Revision of the genus Thinophihis Wahlberg (Diptera: Dolichopodidae) from Singapore and adjacent regions: A long term study with a prudent reconciliation of a genetic to a classic morphological approach**. *Raffles Bulletin of Zoology* 2018, **66**:413-473.
44. Tang CF, Grootaert P, Yang D: **Protomedetera, a new genus from the Oriental and Australasian realms (Diptera, Dolichopodidae, Medeterinae)**. *ZooKeys* 2018, **743**:137-151.
45. Tang CF, Yang D, Grootaert P: **Revision of the genus Lichtwardtia Enderlein in Southeast Asia, a tale of highly diverse male terminalia (Diptera, Dolichopodidae)**. *ZooKeys* 2018, **798**:63-107.
46. Grootaert P: **Species turnover between the northern and southern part of the South China Sea in the Elaphropeza Macquart mangrove fly communities of Hong Kong and Singapore (Insecta: Diptera: Hybotidae)**. *European Journal of Taxonomy* 2019, **554**:1-27.
47. Samoh AC, Satasook C, Grootaert P: **NGS-barcodes, haplotype networks combined to external morphology help to identify new species in the mangrove genus Ngirhaphium Evenhuis & Grootaert, 2002 (Diptera: Dolichopodidae: Rhaphiinae) in Southeast Asia**. *Raffles Bulletin of Zoology* 2019, **67**:640-659.
48. Ismay B, Ang YC: **First records of Pseudogaurax Malloch 1915 (Diptera: Chloropidae) from Singapore, with the description of two new species discovered with NGS barcodes.** *Raffles Bulletin of Zoology* 2019, **67**:412-420.
49. Wang WY, Yamada A, Eguchi K: **First discovery of the mangrove ant Pheidole sexspinosa Mayr, 1870 (Formicidae: Myrmicinae) from the Oriental region, with redescriptions of the worker, queen and male**. *Raffles Bulletin of Zoology* 2018, **66**:652-663.
50. Wang WY, Yong GWJ, Jaitrong W: **The ant genus Rhopalomastix (Hymenoptera: Formicidae: Myrmicinae) in Southeast Asia, with descriptions of four new species from Singapore based on morphology and DNA barcoding**. *Zootaxa* 2018, **4532**(3):301-340.
51. Wang WY, Yamada A, Yamane S: **Maritime trap-jaw ants (Hymenoptera, Formicidae, Ponerinae) of the Indo-Australian region - redescription of Odontomachus malignus Smith and description of a related new species from Singapore, including first descriptions of males**. *ZooKeys* 2020, **915**:137-174.
52. Wick RR: **Performance of neural network basecalling tools for Oxford Nanopore sequencing**. *Genome biology* 2019, **20**:129.
53. Silvestre-Ryan J, Holmes I: **Pair consensus decoding improves accuracy of neural network basecallers for nanopore sequencing**. *Genome biology* 2021, **22**:38.
54. Chang JJM, Ip YCA, Ng CSL, Huang D: **Takeaways from Mobile DNA Barcoding with BentoLab and MinION**. *Genes* 2020, **11**:1121.
55. Vereecke N, Bokma J, Haesebrouck F, Nauwynck H, Boyen F, Pardon B, Theuns S: **High quality genome assemblies of Mycoplasma bovis using a taxon-specific Bonito basecaller for MinION and Flongle long-read nanopore sequencing**. *BMC Bioinformatics* 2020, **21**:517.
56. Sahlin K, Lim MCW, Prost S: **NGSpeciesID: DNA barcode and amplicon consensus generation from long-read sequencing data**. *Ecology and evolution* 2021, **11**(3):1392-1398.
57. Ho JKI, Puniamoorthy J, Srivathsan A, Meier R: **MinION sequencing of seafood in Singapore reveals creatively labelled flatfishes, confused roe, pig DNA in squid balls, and phantom crustaceans**. *Food Control* 2020, **112**:107144.

1035 58. Cranston PS, Ang A, Heyzer A, Lim RBH, Wong WH, Woodford JM, Meier R: **The nuisance**
1036 **midges (Diptera: Chironomidae) of Singapore's Pandan and Bedok reservoirs**. *Raffles*
1037 *Bulletin of Zoology* 2013, **61**:779-793.
1038 59. Xu Z, Mai Y, Liu D, He W, Lin X, Xu C, Zhang L, Meng X, Mafofo J, Zaher WA *et al*: **Fast-Bonito:**
1039 **A Faster Basecaller for Nanopore Sequencing**. *BioRxiv*
1040 2020:doi:10.1101/2020.1110.1108.318535.
1041 60. Chang JJM, Ip YCA, Bauman AG, Huang D: **MinION-in-ARMS: Nanopore Sequencing to**
1042 **Expedite Barcoding of Specimen-Rich Macrofaunal Samples From Autonomous Reef**
1043 **Monitoring Structures**. *Frontiers in Marine Science* 2020, **7**:448.
1044 61. Menegon M, Cantaloni C, Rodriguez-Prieto A, Centomo C, Abdelfattah A, Rossato M,
1045 Bernardi M, Xumerle L, Loader S, Delledonne M: **On site DNA barcoding by nanopore**
1046 **sequencing**. *PlOS One* 2017, **12**(10):e0184741.
1047 62. Wurzbacher C, Larsson E, Bengtsson-Palme J, den Wyngaert SV, Svantesson S, Kristiansson E,
1048 Kagami M, Nilsson RH: **Introducing ribosomal tandem repeat barcoding for fungi.** *Molecular*
1049 *ecology resources* 2018, **19**(1):118-127.
1050 63. Krehenwinkel H, Pomerantz A, Henderson JB, Kennedy SR, Lim JY, Swamy V, Shoobridge JD,
1051 Graham N, Patel NH, Gillespie RG *et al*: **Nanopore sequencing of long ribosomal DNA**
1052 **amplicons enables portable and simple biodiversity assessments with high phylogenetic**
1053 **resolution across broad taxonomic scale**. *GigaScience* 2019.
1054 64. Maestri S, Cosetino E, Paterno M, Freitag H, Garces JM, Marcolungo L, Alfano M, Njunjić I,
1055 Schilthuizen M, Slik F *et al*: **A Rapid and Accurate MinION-Based Workflow for Tracking**
1056 **Species Biodiversity in the Field**. *Genes* 2019, **10**(6):468.
1057 65. Knot IE, Zouganelis GD, Weedall GD, Wich SA, Rae R: **DNA Barcoding of Nematodes Using**
1058 **the MinION**. *Frontiers in Ecology and Evolution* 2020, **8**:100.
1059 66. Seah A, Lim MCW, McAloose D, Prost S, Seimon TA: **MinION-Based DNA Barcoding of**
1060 **Preserved and Non-Invasively Collected Wildlife Samples**. *Genes* 2020, **11**(4):445.
1061 67. Daily J: **Parasail: SIMD C library for global, semi-global, and local pairwise sequence**
1062 **alignments**. *BMC Bioinformatics* 2016, **17**:81.
1063 68. Sahlin K, Medvedev P: **De Novo Clustering of Long-Read Transcriptome Data Using a**
1064 **Greedy, Quality Value-Based Algorithm**. *Journal of Computational Biology* 2020, **27**(4):472-
1065 484.
1066 69. Elbrecht V, Leese F: **Can DNA-Based Ecosystem Assessments Quantify Species Abundance?**
1067 **Testing Primer Bias and Biomass—Sequence Relationships with an Innovative**
1068 **Metabarcoding Protocol**. *PLoS One* 2015, **10**(7):e0130324.
1069 70. Buchner D, Beermann AJ, Leese F, Weiss M: **Cooking small and large portions of**
1070 **"biodiversity-soup": Miniaturized DNA metabarcoding PCRs perform as good as large-**
1071 **volume PCRs**. *Ecology and evolution* 2021, **11**(13):9092-9099.
1072 71. Hickerson MJ, Meyer CP, Moritz: **DNA Barcoding Will Often Fail to Discover New Animal**
1073 **Species over Broad Parameter Space**. *Systematic Biology* 2006, **55**(5):729-739.
1074 72. Valan M, Makonyi K, Maki A, Vondráček D, Ronquist F: **Automated Taxonomic Identification**
1075 **of Insects with Expert-Level Accuracy Using Effective Feature Transfer from Convolutional**
1076 **Networks**. *Systematic Biology* 2019, **68**(6):876-895.
1077 73. Wührl I, Pylatiuk C, Giersch M, Lapp F, von Rintelen T, Balke M, Schmidt S, Cerretti P, Meier
1078 R: **DiversityScanner: Robotic discovery of small invertebrates with machine learning**
1079 **methods**. *BioRxiv* 2021, **doi**:10.1101/2021.1105.1117.444523.
1080 74. Leray M, Yang JY, Meyer CP, Mills SC, Agudelo N, Ranwez V, Boehm JT, Machida RJ: **A new**
1081 **versatile primer set targeting a short fragment of the mitochondrial COI region for**
1082 **metabarcoding metazoan diversity: application for characterizing coral reef fish gut**
1083 **contents**. *Front Zool* 2013, **10**:34.

1084 75. Folmer O, Black M, Hoeh W, Lutz R, Vrijenhoek R: **DNA primers for amplification of**
1085 **mitochondrial cytochrome c oxidase I from diverse metazoan invertebrates**. *Mol Marina*
1086 *Biol Technol* 1994, **3**(5):294-299.
1087 76. Geller J, Meyer C, Parker M, Hawk H: **Redesign of PCR primers for mitochondrial**
1088 **cytochrome c oxidase subunit I for marine invertebrates and application in all-taxa biotic**
1089 **surveys**. *Molecular ecology resources* 2013, **13**(5):851-861.
1090 77. Truett G, Heeger P, Mynatt R, Truett A, Walker J, Warman MJB: **Preparation of PCR-quality**
1091 **mouse genomic DNA with hot sodium hydroxide and tris (HotSHOT)**. *Biotechniques* 2000,
1092 **29**(1):52-54.
1093 78. Thongjued K, Chotigeat S, Bumrungsri P, Thahakiatkrai P, Kitpipit T: **A new cost-effective and**
1094 **fast direct PCR protocol for insects based on PBS buffer**. *Molecular ecology resources* 2019,
1095 **19**(3):691-701.
1096 79. Arribas P, Andújar C, Hopkins K, Shepherd M, Vogler AP: **Metabarcoding and mitochondrial**
1097 **metagenomics of endogean arthropods to unveil the mesofauna of the soil**. *Methods in*
1098 *Ecology and Evolution* 2016, **7**(9):1071-1081.
1099 80. Crampton-Platt A, Yu DW, Zhou X, Vogler AP: **Mitochondrial metagenomics: letting the**
1100 **genes out of the bottle**. *GigaScience* 2016, **5**(1):s13742-13016-10120-y.
1101 81. Zizka VM, Elbrecht V, Macher JN, Leese F: **Assessing the influence of sample tagging and**
1102 **library preparation on DNA metabarcodin**. *Molecular ecology resources* 2019, **19**(4):893-
1103 899.
1104 82. Valentini A, Miquel C, Nawaz MA, Bellemain E, Coissac E, Pompanon F, Gielly L, Cruaud C,
1105 Nascetti G, Wincker P *et al*: **New perspectives in diet analysis based on DNA barcoding and**
1106 **parallel pyrosequencing: the trnL approach**. *Molecular ecology resources* 2009, **9**(1):51-60.
1107 83. Rohland N, Reich D: **Cost-effective, high-throughput DNA sequencing libraries for**
1108 **multiplexed target capture**. *Genome research* 2012, **22**:939-946.
1109 84. Frith MC, Mitsuhashi S, Katoh K: **lamassemble: Multiple Alignment and Consensus**
1110 **Sequence of Long Reads**. In: *Multiple Sequence Alignment.* Edited by Katoh K. New York:
1111 Humana; 2020: 135-145.
1112 85. Hamada MY, Ono Y, Asai K, Frith MC: **Training alignment parameters for arbitrary**
1113 **sequencers with LAST-TRAIN**. *Bioinformatics* 2017, **33**(6):926-928.

1114