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]

[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

Abstract

DNA barcodes are a useful tool for discovering, understanding, and monitoring biodiversity. This is critical at a time when biodiversity loss is a major problem for many countries. However, widespread adoption of barcoding programs requires the process to be cost-effective and simple to apply. 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 real-time sequencer MinION. We first describe cost-effective and rapid procedures in a comprehensive workflow for obtaining tagged amplicons. We then demonstrate how a portable MinION device can be used for real-time sequencing of tagged amplicons in many settings (field stations, biodiversity labs, citizen science labs, schools). 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. We also provide amplicon coverage recommendations that are based on several runs of MinION flow cells (R10.3) involving >24,000 specimen barcodes, which suggest that each run can generate >10,000 barcodes. Additionally, we present a novel software, ONTbarcoder, that overcomes the bioinformatics challenges posed by the sequencing errors of MinION reads. This software is compatible with Windows10, Macintosh, and Linux, has a graphical user interface (GUI), and can generate thousands of barcodes on a standard laptop within hours based on two input files (FASTQ, demultiplexing file). Next, we document that MinION barcodes are virtually identical to Sanger and Illumina barcodes for the same specimens (>99.99%). Lastly, we demonstrate how rapidly MinION data have improved by comparing the performance of sequential flow cell generations. We overall assert 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. Biodiversity loss is threatening the planet and the use of MinION barcodes will help with enabling an army of researchers and citizen scientists, which is necessary for effective biodiversity discovery and monitoring.

1. Background

DNA sequences have been used for identification and taxonomic purposes for decades (Hebert, Cywinska et al. 2003, Tautz, Arctander et al. 2003, Meier 2008), but for most of this time been akin to mobile phones in the 1990s: of limited value due to sparse signal coverage and high cost. Obtaining barcodes was problematic due largely to the complicated and expensive procedures on which it relied. Some of these problems have since been addressed by, for example, developing effective DNA extraction protocols and optimizing Sanger sequencing procedures (Ivanova, Dewaard et al. 2006, Ivanova, Borisenko et al. 2009). These improvements enabled the establishment of a centralized barcoding facility in 2006. After 15 years and the investment of >200 million USD, ca. 8.3 million barcodes are available for searches on BOLD Systems, but only 2.2 million of these are in the public domain (http://boldsystems.org/index.php/IDS_OpenIdEngine). Combined with barcodes from NCBI GenBank, they are now a valuable resource to the global biodiversity community. However, the cost of barcodes has remained high (http://ccdb.ca/pricing/) and the current approach that requires sending specimens from all over the world to one center and then back to the country of origin interferes with real-time biodiversity monitoring and specimen accessibility. We would therefore argue that access to barcodes has to be decentralized and we believe that the best strategy for achieving this goal is by applying a technique that is known as “innovation through subtraction” in engineering. It usually delivers simplified and often more cost-effective solutions by challenging conventions. Fortunately, DNA barcoding is imminently suitable for this innovation strategy because the established methods have numerous legacy issues. Indeed, we here show that the amplification and sequencing of a short mitochondrial COI fragment can be efficiently performed anywhere.

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 (Stork, McBroom et al. 2015) and a new species is discovered with every 50th specimen that is processed, species discovery with barcodes will require the sequencing of 500 million specimens (Yeo, Srivathsan et al. 2020). 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 (World Economic Forum 2020) and Swiss Re estimates that 20% of all countries face ecosystem collapse as biodiversity declines (Swiss Re 2020). Biodiversity loss is no longer just an academic concern; it is now a major threat to human communities and the health of the planet. This also implies that biodiversity discovery and monitoring have to be accomplished at completely different scales than in the past. The old approaches thus need rethinking because all countries need distributional and abundance information to develop effective conservation strategies and policies. In addition, they need information on how species interact with each other and the environment. Many of these biodiversity monitoring and environmental management activities have to focus on terrestrial invertebrates, whose biomass surpasses that of all terrestrial vertebrates combined (Bar-On, Phillips et al. 2018) and who occupy a broad range of ecological guilds. Many of these invertebrate clades are extremely specimen- and species-rich which means that monitoring should be locally conducted to allow for rapid turnaround times. This also means that it will be important to have simple and cost-effective procedures that can be implemented anywhere by stakeholders with very different scientific and skill backgrounds.

DNA barcoding was proposed at a time when biodiversity loss was not on the radar of economists. Instead, barcodes were initially intended as an identification tool for biologists (Hebert, Cywinska et al. 2003). Thus, most projects focused on taxa with a large following in biology (e.g., birds, fish, butterflies) (Kwong, Srivathsan et al. 2012). However, this also meant that these projects only covered a small proportion of the terrestrial animal biomass (Bar-On, Phillips et al. 2018) and species-level diversity (Groombridge 1992). Yet, despite targeting taxa with well-understood diversity, the projects struggled with covering >75% of the described species in these groups (Kwong, Srivathsan et al. 2012). When the pilot barcoding projects ran out of material from identified specimens, they started targeting unidentified specimens; i.e., DNA barcoding morphed into a technique that was used for biodiversity discovery (“dark taxa”: (Page 2011, Kwong, Srivathsan et al. 2012). This shift towards biodiversity discovery was gradual and incomplete because the projects used a “hybrid approach” that started with subsampling or sorting specimens to “morphospecies” before barcoding representatives of each morphospecies/sample (e.g., (Barrett and Hebert 2005, Hendrich, Pons et al. 2010, Hebert, DeWaard et al. 2013, Ng'endo, Osiemo et al. 2013, Hebert, Ratnasingham et al. 2016, Thormann, Ahrens et al. 2016, Knox, Hogg et al. 2020). This is problematic, as morphospecies sorting is known to be labour-intensive and of unpredictable quality because it is heavily dependent on the taxonomic expertise of the sorters (Krell 2004, Stribling, Pavlik et al. 2008). Thus, such hybrid approaches are of limited value for obtaining reliable quantitative data on biodiversity, but were adopted as a compromise owing to the prohibitive cost of barcoding. The logical alternative is to barcode all specimens and then group them into putative species based on sequence information. The stability and reliability of these groupings can then be evaluated by applying different species delimitation algorithms and by testing the units using other data (e.g., morphology, nuclear markers). Such a “reverse workflow” (Wang, Srivathsan et al. 2018), where every specimen is barcoded as the initial pre-sorting step, yields quantitative data and corroborated species-level units. However, the reverse workflow 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 lowered sequencing costs dramatically (Shokralla, Spall et al. 2012, Shokralla, Porter et al. 2015, Meier, Wong et al. 2016, Hebert, Braukmann et al. 2018, Krehenwinkel, Kennedy et al. 2018, Srivathsan, Baloglu et al. 2018, Wang, Srivathsan et al. 2018, Srivathsan, Hartop et al. 2019, Yeo, Srivathsan et al. 2020). Such changes mean that the reverse workflow is now available 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 (Ponder and Lunney 1999, Srivathsan, Hartop et al. 2019). Many of these clades have high spatial species turnover, requiring many localities in each country to be sampled and massive numbers of specimens to be processed (Yeo, Srivathsan et al. 2020). Such intensive processing is best achieved close to the collecting locality to avoid the unnecessary risks, delays and cost from shipping biodiversity 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 and can in turn significantly enhance biodiversity education and appreciation (Pomerantz, Peñafiel et al. 2018, Watsa, Erkenswick et al. 2020). This is especially so when stakeholders not only barcode, but can also image specimens, determine species abundances, and map distributions of newly discovered species. All of which may come from specimens collected in their own backyard.

But can such decentralized biodiversity discovery really be effective? Within the last five years, the laboratory of the corresponding author at the National University of Singapore has barcoded >330,000 specimens. Much of the work was carried out by students and interns and yielded the kind of information that countries now need to initiate holistic biodiversity assessment. Singapore represents a typical urbanized environment in that (1) only charismatic taxa are well known, (2) 90% of its original vegetation cover has been lost, and (3) the country is strongly affected by global warming while depending on its remaining forests and urban vegetation for many ecosystem services. Over the past ten years, we have addressed the knowledge gaps for terrestrial arthropods through a Malaise trap program that eventually covered 107 sites and yielded an estimated 4-5 million specimens (Yeo, Srivathsan et al. 2020). After analyzing the first >200,000 barcoded specimens for selected taxa representing different ecological guilds, the alpha and beta diversity of

Singapore's arthropod fauna could be analyzed based on ~8,000 putative species collected across 6 habitat types (mangroves, rainforests, swamp forests, disturbed secondary urban forests, dry coastal forests, freshwater swamps). This revealed that some habitats were unexpectedly species-rich and harboured very unique faunas (e.g., mangroves). Barcodes were also instrumental in revealing that even small remnants of a natural habitat can remain resistant to the invasion of species from neighbouring man-made habitats (Baloğlu, Clews et al. 2018) and in helping 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 (Yeo, Puniamoorthy et al. 2018). This large and comprehensive local barcode database also facilitated species interaction research and biodiversity surveys based on eDNA (Lim, Tay et al. 2016, Srivathsan, Nagarajan et al. 2019). In order to foster biodiversity appreciation, many images of the newly discovered species and their species interactions were placed on the "Biodiversity of Singapore" (BOS) website which now features >15,000 species (https://singapore.biodiversity.online/).

In addition to such contributions to biodiversity knowledge, the widespread application of the reverse workflow has proved a boon for integrative taxonomy, facilitating modern taxonomy in many ways. Firstly, taxonomic experts do not have to spend time on time-consuming morphospecies sorting involving thousands of specimens, and can instead focus on establishing whether putative species delimited with DNA barcodes are valid. This is a necessary step before species description given that DNA barcodes are far from being an infallible tool for species delimitation and often yield different putative species numbers and compositions when analyzed with different tools (Kekkonen, Mutanen et al. 2015, Ahrens, Fujisawa et al. 2016, Yeo, Srivathsan et al. 2020). Secondly, all specimens that are studied have associated sequence information which identifies which species are closely related and should be compared. This is particularly advantageous when additional specimens are sequenced at a later date as they can immediately get associated to a species for comparative work.

In Singapore, many of the putative species are featured on the BOS website where they are discovered by taxonomic specialists who borrow material for follow-up study. The use of the reverse workflow in Singapore has thus led to an acceleration of biodiversity discovery and description, with dozens of new species already described and the descriptions of another 150 species being finalized (Grootaert 2018, Tang, Grootaert et al. 2018, Tang, Yang et al. 2018, Wang, Yamada et al. 2018, Wang, Yong et al. 2018, Grootaert 2019, Ismay and Ang 2019, Samoh, Satasook et al. 2019, Wang, Yamada et al. 2020).

2. Methods for the democratization of DNA barcoding through simplification

Barcoding a metazoan specimen requires the successful completion of three steps: (1) obtaining DNA template, (2) amplifying *COI* via PCR, and (3) sequencing the *COI* amplicon. Most scientists 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 and simplifying complicated, time-consuming, and expensive procedures. Overall, it is a typical implementation of “innovation through subtraction”. Note that this unlearning is of critical importance for the democratization of biodiversity discovery with DNA barcodes and is particularly vital for boosting biodiversity research where it is most needed: in biodiverse countries with limited science funding.

In this section, we first briefly summarize commonly used procedures for DNA extraction, PCR, and sequencing. For each step we then describe how the procedures can be simplified. Note that all techniques have been extensively tested in our lab, primarily on invertebrates preserved in ethanol for species discovery. Regarding sequencing, we briefly introduce four methods, but focus on MinION sequencing because we recently tested the latest flow cells (R10.3 and Flongle). Both performed very well and we here argue that they are particularly suitable as the default sequencing option for decentralized biodiversity discovery. The results of these tests and a new software package for MinION barcoding are presented in the third part of this paper.

Methods for step 1: Obtaining DNA template

Most biologists learn that DNA extraction requires tissue digestion with a proteinase, purification of the DNA, and finally the elution of DNA. This approach is slow and expensive because it frequently involves kits and consumables that are designed for obtaining the kind of high-quality DNA that is needed for amplifying "difficult" genes (e.g., long, single-copy nuclear markers). However *COI* is a mitochondrial gene and thus naturally enriched. Indeed, the mitochondrial genome is tiny (16 kbp) and yet usually contributing 0.5-5% of the DNA in a genomic extraction (Arribas, Andújar et al. 2016, Crampton-Platt, Yu et al. 2016). Furthermore, barcoding requires only the amplification of one short marker (<700 bp) so that not much DNA template is needed. This allows for using the following simplified procedures that are designed for specimens containing DNA template of reasonable quality.

Simplified DNA "extraction": Obtaining template for DNA barcoding need not take more than 20 minutes, does not require DNA purification, and costs essentially nothing. The cheapest, but not necessarily fastest, method is "directPCR"; i.e., deliberately "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). This method is very fast and effective for small specimens lacking thick cuticle or skin (Wong, Tay et al. 2014) and works particularly well for many abundant aquatic invertebrates such as chironomid midges and larvae. Larger specimens require the use of body parts (leg or antenna: Wong, Tay et al. (2014)). Such dissections tend to be labour-intensive if large numbers of specimens must be processed, but it is a good method for small numbers of samples or in barcoding experiments that are carried out in poorly equipped labs. Note that the whole body or body part that is used for directPCR can be recovered after amplification, although soft-bodied animals may become transparent.

An alternative to directPCR is buffer-based DNA extraction. This method is also essentially cost-free because it involves alkaline buffers that are inexpensive, usually available in molecular labs (e.g., PBS), or can be prepared easily (HotSHOT (Truett, Heeger et al. 2000, Thongjued, Chotigeat et al. 2019)). Our preferred method is extraction with HotSHOT, which we have used for barcoding >50,000 arthropods. We use 10-15 µL HotSHOT per specimen. Small specimens are submerged within the well of a microplate while larger specimens are placed head-first into the well. DNA is obtained within 20 minutes in a thermocycler via two heating steps (Truett, Heeger et al. 2000). After neutralization, >20 µl of template is available for amplifying *COI* and the voucher can be recovered. Note that HotSHOT extraction leaves most of the DNA in the specimen untouched and more high quality DNA can subsequently be extracted from the same specimen. An alternative to obtaining DNA via lab buffers is the use of commercial DNA extraction buffers (Kranzfelder, Ekrem et al. 2016). These buffers have a longer shelf life, and are good alternatives for users who only occasionally barcode moderate numbers of specimens. In the past, we have used QuickExtract (Srivathsan, Hartop et al. 2019) and found that 10 µl is sufficient for obtaining DNA template from most insect specimens. In summary, obtaining DNA templates for barcoding is fast and straightforward and most published barcoding studies greatly overcomplicate this step. It should be noted however, that all DNA extraction methods require the removal of excess ethanol from specimens prior to extraction (e.g., by placing the specimen on tissue paper or replacing ethanol with water prior to specimen processing) and that the DNA extracts obtained with such methods have a short shelf-life even in a freezer.

Methods for step 2: amplifying *COI* via PCR. Like procedures for DNA extraction, most PCR recipes and reagents are optimized to work for a wide variety of genes and not just for a gene like the *COI* barcode that is naturally enriched, has a large number of known primers, and is fairly short. Standard PCR recipes can therefore be simplified. However, the use of sequencing technologies such as Illumina, PacBio, or Oxford Nanopore Technologies introduces one complication: The amplicons have to be “tagged” (or “indexed”/“barcoded”). This is necessary because modern sequencing instruments sequence a pool of amplicons simultaneously instead of processing one amplicon at a time (as in Sanger sequencing). Tags are short DNA sequences that are attached to the 5’ end of the amplicon and can then be used as a specimen identifier. This allows for the assignment of each read obtained during sequencing to the amplicon obtained for a specific specimen (“demultiplexing”). Numerous tagging techniques have been described in the literature, but these, too, can be greatly simplified for DNA barcoding.

Simplified techniques for obtaining tagged amplicons

Published protocols tend to have four issues that increase workload and/or inflate cost, while a fifth issue only affects amplicon tagging:

- *Issue 1: expensive polymerases or master mixes.*
  These often utilize high-fidelity polymerases that are designed for amplifying low copy-number nuclear genes based on low-concentration template but rarely make a difference when amplifying *COI*. Indeed, even home-made polymerases can be used for barcoding. This is important because high import taxes interfere with biodiversity discovery in many biodiverse countries.
- *Issue 2: indiscriminate use of single-use consumables.*
  Disposable products increase costs and damage the environment. Most biodiversity samples are obtained under “unclean conditions” that increase the chance for cross-specimen contamination long before specimens reach the lab (e.g., thousands of specimens rubbing against each other in sample containers and in the same preservation fluid). Yet numerous studies have shown that the DNA from specimens exposed to such conditions will usually outcompete contaminant DNA that is likely to occur at much lower concentrations. Similarly, the probability that a washed/flushed and

autoclaved microplate or pipette tip retains enough viable contaminant DNA to successfully outcompete the template DNA is extremely low. Indeed, we have repeatedly tried and failed to amplify *COI* using reused plastic consumables and water as template. That it is safe to reuse some consumables is again good news for biodiversity discovery under severe financial constraints. Note, however, that we do not recommend the repeat use of consumables for handling stock chemicals such as primers and sequencing reagents.

- *Issue 3: large PCR volumes (25-50 µl).*

  Pools of tagged amplicons comprise hundreds or thousands of products and there is typically more than enough DNA for preparing a library. Accordingly, even small PCR volumes of 10-15 µl are sufficient, thereby reducing consumable costs for PCR to nearly half when compared to standard volumes of 25-50 µl.

- *Issue 4: using gel electrophoresis for checking amplification success of each PCR product.*

  This time-consuming step is only justified when Sanger sequencing is used or when high-priority specimens are barcoded. It is not necessary when barcoding large numbers of specimens with modern sequencing technologies, because failed amplicons do not add to the sequencing cost. Furthermore, specimens that failed to yield barcodes during the first sequencing run can be re-sequence or re-amplified and then added to subsequent sequencing runs. We thus only use gel electrophoresis to check a small number of reactions per microplate (N=8-12, including the negative control) in order to make sure that there was no plate-wide failure.

The fifth issue requires more elaboration and concerns how to efficiently tag amplicons so that the sequencing reads can be traced back to a specific specimen. We tag our amplicons via a single PCR reaction (Meier, Wong et al. 2016) using primers synthesized with the tag at the 5' end because it is simpler than the dual-PCR tagging strategy dominating the literature. The latter has numerous disadvantages when applied to one gene: it doubles the cost by requiring two rounds of PCR, is more labour intensive, increases the risk for PCR errors by requiring more cycles, and requires clean-up of every PCR product after the first round of amplification. In contrast, tagging via a single PCR is simple and costs the same as any gene amplification. It is here described for a microplate with 96 templates, but the protocol can be adapted to the use of strip tubes or half-plates. What is needed is a 96-well primer plate where each well contains a reverse primer that has a different tag. This primer plate can yield 96 unique combinations of primers once the 96 reverse primers are combined with the same tagged forward primer (1 identically tagged 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 below).

Assigning tag combinations is also straightforward. For each plate with 96 PCR reactions, add one tagged f-primer to a tube with the master mix of routine PCR reagents (Taq DNA polymerase, buffer and dNTPs) for the plate. Then dispense the “f-primed” master mix into the 96-wells. Afterwards, 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 the tracking of tag combinations simple because each PCR plate has its own tagged f-primer to record, while the r-primer is consistently tied to well position. Each plate has a negative control to ensure that no widespread contamination has occurred. The tagging information for each plate is recorded in the demultiplexing file that is later used to demultiplex the reads obtained during sequencing.

Some users may worry that the purchase of so many primers is expensive, but one must keep in mind that the amount of primer used per PCR reaction is constant. Therefore, single PCR-tagging only means a greater upfront investment, but costs half that of dual PCR-tagging. However, ordering all primers at once does mean that one must be much more careful about avoiding primer degeneration and contamination as the stock will last longer. This is because 1 nmol of primer can be used for ~50 reactions (=microplates). Primer stock should be stored at -80°C and the number of freeze-thaw cycles should be kept low (<10). This means that upon receipt of the primer stock, it should be immediately aliquoted into plates/tubes holding only enough primer for rapid use. For fieldwork, one should only bring enough dissolved primer for the necessary experiments, or rely on lyophilised reagents.

The choice of tag length is determined by three factors. Longer tags reduce PCR success rates (Srivathsan, Hartop et al. 2019) while they increase the proportion of reads that can be assigned to a specific specimen (demultiplexing rate). Designing tags is not straightforward because they must remain sufficiently distinct (>4bp from each other including insertions/deletions) while avoiding homopolymers. We include the 13 bp tagged primers that we use for MinION based barcoding in supplementary materials. Note, however, that we here also re-sequenced an older amplicon pool that used 12 bp tags (Srivathsan, Baloglu et al. 2018).

Methods for step 3: Amplicon sequencing. The use of the PCR techniques described so far should keep the cost for a tagged barcode amplicon to 0.05-0.10 USD as long as the user buys cost-effective consumables. What comes next is the purification of the amplicons via the removal of unused PCR reagents and an assessment/adjustment of DNA concentration. This only has to be done for each amplicon separately when Sanger sequencing is used. The sequencing alternatives to Sanger sequencing are Oxford Nanopore Technologies (ONT) (e.g., MinION), Illumina (e.g. NovaSeq), and PacBio (e.g., Sequel) for which large-scale sequencing protocols have been described (Hebert, Braukmann et al. 2018, Wang, Srivathsan et al. 2018, Srivathsan, Hartop et al. 2019). Users can select the sequencing option that best suit their needs. Five criteria matter: (1) Scaling (ability to accommodate projects of different scales), (2) turnaround times, (3) cost, (4) amplicon length and (5) sequencing error rate. For example, Sanger sequencing has fast turnaround times but higher sequencing costs per amplicon ($3-4 USD). This is the only method where cost scales linearly with the number of amplicons that need sequencing, while the other sequencing techniques are fundamentally different in that each run has two fixed costs that stay the same regardless of whether only a few or the maximum number of amplicons for the respective flow cells are sequenced. The first such cost is “library preparation” (getting amplicons ready for sequencing) and the second is the flow cell that is used for sequencing.

The MinION Flongle has the lowest run cost among the 2nd and 3rd generation sequencing techniques (library and flow cell: ca. $140 USD), which we show in this paper to have sufficient capacity for ca. 250 barcode amplicons. The turnaround time is fast, so the MinION Flongle is arguably the best sequencing option for small barcoding projects that require the sequencing > 50 barcodes. Full MinION flow cells also have fast turnaround times, but the minimum run cost is ca. 1000 USD, so this option only becomes more cost-effective than Flongle when >1800 amplicons are sequenced. As shown later, one regular MinION flow cell can comfortably sequence 10,000 amplicons. This is a similar volume to what has been described for PacBio (Sequel) (Hebert, Braukmann et al. 2018), but the high instrument cost for PacBio means that sequencing usually has to be outsourced, leading to longer wait times. By far the most cost-effective sequencing method for barcodes is Illumina’s NovaSeq sequencing. The fixed costs for library and lanes are high (3000-4000 USD), but each flow cell yields 400 million reads which can comfortably sequence 400,000 barcodes at a cost of < $0.01USD per barcode. This extreme capacity means that all publicly available barcodes in BOLD Systems could have been sequenced on just five NovaSeq flow cells for ~20,000 USD. However, Illumina sequencing can only be used for mini-barcodes of up to 400 bp length (using 250bp PE sequencing using SP flow cell). The full-length *COI* barcode (658 bp) can only be retrieved by sequencing both halves separately. Note that while Illumina barcodes are shorter than “full-length” barcodes, a recently published study found no evidence that minibarcodes have a negative impact on species delimitation or identification as long as the mini-barcode is >250bp in length (Yeo, Srivathsan et al. 2020).

Simplified techniques for sequencing tagged amplicons: Modern sequencing technologies are used to sequence amplicon pools. To obtain such a pool, it is sufficient to combine only 1 µl per PCR product. The pool can be cleaned using several PCR clean-up methods. We generally use 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 (Rohland and Reich 2012) are also viable options. We recommend the use of a 0.5X ratio for Ampure beads for barcodes longer than 300 bp 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. Amplicon pools containing large numbers of amplicons usually require multiple rounds of clean-up, but only a small subset of the entire pool needs to be purified because most library preparation kits require only small amounts of DNA. Note that the success of the clean-up procedures should be verified with gel electrophoresis, which should yield 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. Most laboratories use a Qubit, but less precise techniques may also be suitable.

Obtaining a cleaned amplicon pool according to the outlined protocol is not time consuming. However, many studies retain “old Sanger sequencing habits” although they use modern sequencing technologies. For example, they use gel electrophoresis for each PCR reaction to test whether an amplicon has been obtained and then clean and measure all amplicons one at a time for normalization (often with very expensive techniques: Ampure beads: (Maestri, Cosetino et al. 2019); TapeStation, BioAnalyzer, Qubit: (Seah, Lim et al. 2020)). The goal is to obtain a pool of amplicons where each has equal representation. Such a pool indeed has the attractive property of each amplicon yielding a similar number of sequencing reads regardless of the initial yield during PCR. However, reads are cheap while individual clean-ups and measurements are expensive. A more cost-effective approach is equalizing amplicon coverage via resequencing. One can first sequence a “raw” amplicon pool with moderate coverage. Afterward, the number of reads in each specimen-specific read bins can be determined. This reveals the weak amplicons that can then be re-sequenced in order to obtain higher coverage (see (Srivathsan, Hartop et al. 2019). Yet another alternative is to use gel electrophoresis for a handful of products per PCR microplate to classify entire plates as being “strong”, “weak”, or “largely failed”. Then three amplicon pools can be prepared, and the DNA contribution of each pool can be adjusted to accurately reflect the number amplicons in each pool. For example, a pool of 500 amplicons from “weak” plates may have only half the DNA concentration of a pool of 500 amplicons from “strong” plates. For the final pool, the “weak pool” should contribute twice the volume of the “strong” pool.

3. Testing MinION barcoding with new flow cells (R10.3, Flongle) and high-accuracy basecalling

Oxford Nanopore Technologies (ONT) instruments sequence DNA by passing single-stranded DNA through a nanopore. This creates current fluctuations which can be measured and translated into a DNA sequence via basecalling (Wick 2019). The sequencing devices are small and inexpensive, but the read accuracy is only moderate (85% - 95%) (Wick 2019, Silvestre-Ryan and Holmes 2021). This means that many reads for the same amplicon are needed to reconstruct the amplicon sequence via specialized bioinformatics pipelines. The nanopores used for sequencing are arranged on flow cells, with new flow cell chemistries and basecalling software regularly released. Recently, three significant changes occurred which motivated our new test of MinION barcoding. Firstly, ONT released a flow cell (Flongle) that uses the currently most widely used chemistry (R9.4), but only has 126 pores (126 channels) instead of the customary 2048 pores (512 channels) of a full MinION flow cell. We were interested in Flongle because it looked promising for small barcoding projects that needed quick turnaround times. For them, currently only Sanger sequencing makes financial sense. Secondly, ONT also released new flow cell chemistry (R10.3). The new flow cells have nanopores that have a dual reader-head instead of the single head in R9.4. Dual reading has altered the read error profile by giving better resolution to homopolymers and improving consensus accuracy (Chang, Ip et al. 2020, Vereecke, Bokma et al. 2020). Lastly, ONT released high accuracy (HAC) basecalling which promises more accurate sequencing reads but also affects existing bioinformatics pipelines. HAC basecalling using R10.3 flow cell has been shown to be promising for DNA barcoding, but the test was based only on ca. 100 barcodes (Chang, Ip et al. 2020).

Library preparation. Most of the wet laboratory methods used for the flow cell tests in this manuscript are summarized in Table 1. Library preparation was based on 200 ng of DNA for the full MinION flow cells and 100 ng for the Flongle. All libraries were prepared with ligation-based library preparation kits. We generally followed kit instructions, but excluded 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 consisted of 45 µl of DNA, 7 µl of Ultra II End-prep reaction buffer (New England Biolabs), 3 µl of Ultra II End Prep Reaction Buffer (New England Biolabs) and 5 µl of molecular grade water. For the Flongle, only half of the reagents were used to obtain a total volume of 30 µl. We further modified 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 were loaded and sequenced with 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 2), under the high-accuracy model in MinIT taking advantage of its GPU.

Sequencing. We tested MinION barcoding on the new R10.3 and Flongle flow cells for six amplicon pools (Table 1). For two of the pools, *Mixed Diptera* and *Afrotropical Phoridae*, we have comparison barcodes that were obtained with Sanger and Illumina sequencing. Both sequencing technologies have much lower error rates than the 5-15% reported for individual MinION reads (Wick 2019, Silvestre-Ryan and Holmes 2021): individual bases generated using Illumina sequencing overwhelmingly have an accuracy of >99% so that very accurate consensus barcodes can be obtained, while the Sanger barcodes could be carefully edited using manual inspection of chromatograms. These amplicon pools were also used previously for testing earlier versions of MinION flow cells (Srivathsan, Baloglu et al. 2018, Srivathsan, Hartop et al. 2019) (Table 1). We here used these pools to assess the accuracy of barcodes obtained with MinION R10.3. Two additional datasets, *Palaearctic Phoridae (658)* and *Palaearctic Phoridae (313)* were obtained for the same 9,934 specimens of phorids for which we amplified both the “full-length” barcodes (658bp) and mini-barcodes (313bp). These datasets were used to assess the capacity of R10.3 flow cells. The *Mixed Diptera Subsample* and *Chironomidae* datasets test the performance of the Flongle. The *Mixed Diptera Subsample* (N=257) is a subset of the *Mixed Diptera* amplicon pool for which we have Sanger barcodes for comparison. The *Chironomidae* dataset contains sequences for 313 bp mini-barcodes for 191 specimens of Chironomidae that were newly amplified for this study.

Table 1. 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 (see Srivathsan et al., 2018) - Sanger barcodes available	511 (257 mixed Diptera, 254 Dolichopodidae) 17 negatives	658 bp HCO2198, LCO1490 (Folmer et al., 1994)	Extraction Method: 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	Ampure beads (Beckman Coulter)	SQK-LSK110/FLO-MIN111
Afrotropical Phoridae (see Srivathsan et al., 2019) - Illlumina mini-barcodes available	4275 (Phoridae) 45 negatives	658 bp HCO2198, LCO1490 (Folmer et al., 1994)	Extraction Method: QuickExtract PCR Mix: Total volume: 15.16µl Mastermix (CWBio): 10µl 25mM MgCl2: 0.16µl BSA (1mg/ml): 2µl Primer (10µM): 1µl each DNA: 1µl	Sera-Mag beads (GE Healthcare Life Sciences) in PEG	SQK-LSK109/FLO-MIN111
Palaearctic Phoridae (658)	9,934 (Phoridae) 105 negatives	658 bp jgHCO2198, LCO1490 (Folmer et al., 1994, Geller et al. 2013)	Extraction Method: HotSHOT PCR Mix: Total volume: 16µl Mastermix (CWBio): 7µl BSA (1mg/ml): 1µl Primer (10µM): 1µl each DNA: 6 µl	Ampure beads (Beckman Coulter)	SQK-LSK110/FLO-MIN111
Palaearctic Phoridae (313)	9,934 (Phoridae) 104 negatives	313 bp m1COlintF, jgHCO2198 (Leray et al. 2013, Geller et al. 2013)	Extraction Method: HotSHOT PCR Mix: Total volume: 14µl Mastermix (CWBio): 7µl BSA (1mg/ml): 1µl Primer (10µM): 1µl 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 (Leray et al. 2013, Geller et al. 2013)	Extraction Method: HotSHOT PCR Mix: Total volume: 14µl Mastermix (CWBio): 7µl BSA (1mg/ml): 1µl Primer (10µM): 1µl each DNA: 4µl	Ampure beads (Beckman Coulter)	SQK-LSK109/Flongle

Bioinformatics

One of the most significant barriers to widespread barcoding with MinION is the high error rates of ONT reads. In 2018, we developed a bioinformatics pipeline for error correction that was too complex for the average user (Srivathsan, Baloglu et al. 2018, Srivathsan, Hartop et al. 2019). After obtaining data with several R10.3 and new R9.4 flow cells, we initially applied this miniBarcoder pipeline (Srivathsan et al. 2019), but we noticed major improvements in terms of MinION read quality and the total number of raw and demultiplexed reads produced by each flow cell. We briefly also considered alternative pipelines, but they faced one or several of the following problems: they required high read coverage, relied on external sequences, were complex, and/or needed several command line steps and external dependencies that limit cross platform compatibility (Menegon, Cantaloni et al. 2017, Maestri, Cosetino et al. 2019, Seah, Lim et al. 2020, Sahlin, Lim et al. 2021). We therefore decided that it was time to develop a new software package that is suitable for the more widespread use of MinION for biodiversity discovery. We thus wrote “ONTbarcoder”, which compared to other software packages is faster, has a graphical user interface (GUI), and is suitable for all major operating systems (Linux, Mac OS, Windows10); i.e., this software package can help with the democratization of barcoding with MinION.

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. The user is asked to provide three critical pieces of information and two files: (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 user-specified read length threshold are demultiplexed. Technically, the specified length should be that of the amplicon plus both tagged primers, but ONT reads are occasionally too short and we would advise to subtract ca. 20bp or use the barcode length as the read length threshold. Reads that are twice the expected fragment length are split into two parts. Splitting is based on the user given fragment size, primer and tag lengths, and a window size to account for indel errors (default=100 bp).

Once all reads have been prepared for demultiplexing, 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, 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 to “tag variants” that differ by up to 2 bps from the original tag (substitutions/insertions/deletions). Accepting tag variants does not lead to demultiplexing error because all tags differ by >4bp. All reads thus identified as belonging to the same specimen are pooled into the same bin. To increase efficiency, demultiplexing is parallelized and the search space for primers and tags are restricted to user-specified parts of each read.

b. Barcode calling: Barcode calling uses the reads within each specimen-specific bin to reconstruct each 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. Alignment must be efficient in order to obtain high-quality barcodes at reasonable speed for thousands of amplicons. ONTbarcoder delivers speed by using an iterative approach that gradually increases the number of reads (“coverage”) that is used during alignment. However, reconstructing barcodes based on few reads could lead to errors and which are here weeded out by using four rigorous 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 known length of the barcode are used. For example, if the user specified coverage=25x for a 658bp 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”. The barcodes that failed the QC during the “Consensus by Length” stage are often 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 alignments. Such reads differ considerably 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 previously (including the same QC criteria). This improvement step converts many preliminary barcodes 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 errors. ONTbarcoder identifies errors in such a preliminary barcode 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 identifies insertions and deletions in the remaining preliminary barcodes. 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.

Output. ONTbarcoder extensively documents the barcoding results so that users can check the output and potentially modify the barcode calling parameters. For example, it 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. (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. This filtering thresholds were calibrated based on two datasets for which we have Sanger/Illumina barcodes. Note that the file with filtered barcodes also includes the

QC_compliant barcodes. All results discussed in this manuscript are based on filtered barcodes.

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 contamination.

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. 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 MSAs that use MinION reads to form a consensus barcode are constructed in an approach similar to lamassemble (Frith, Mitsuhashi et al. 2020), using parameters optimized for nanopore data by “last-train” (Hamada, Ono et al. 2017) which accounts for strand specific error biases. The MAFFT parameters can be modified in the “parfile” supplied with the software which will help 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. 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 there is a single consensus sequence obtained. 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 therefore includes a second tab that simplifies such comparisons. A set of barcodes is 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 used here to get the final accuracy values presented in Table 3.

4. Performance of flow cells (R10.3, Flongle) and high-accuracy basecalling

The pools used to test the new ONT products contained amplicons for 191 - 9,934 specimens and were run for 15-49 hours (Table 2). The fast5 files were basecalled using Guppy in MinIT under the high accuracy (HAC) model. Basecalling large datasets under HAC is currently still very slow and took 12 days for the *Palaearctic Phoridae (658 bp)* dataset (Table 2). However, the data called with HAC yielded reads that could be demultiplexed well for three of the four R10.3 MinION datasets (= high demultiplexing rates of 30-49%). The exception was the *Palaearctic Phoridae (313 bp)* dataset which demultiplexed poorly (15.5%). Flongle datasets showed overall also lower demultiplexing rates (17-21%).

Table 2. 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%/3,725/121/59/247 Total success rate= 3905/4275 (91.3%)
Palaearctic Phoridae (658 bp, N=9,934)	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= 8365/9934 (84.2%)
Palaearctic Phoridae (313 bp, N=9,934)	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= 8935/9934 (89.9%)
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 used ONTbarcoder to analyze the MinION 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 with either coverage or time on the x-axis.

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. This means that ONTbarcoder selected up to 200 reads from the specimen-specific read bin that had the closest match to the length of the target barcode (i.e., 313 or 658 bp), then produced an MSA and consensus barcode using MAFFT. If the resulting consensus barcode did not satisfy all four QC criteria, ONTbarcoder would select up to 100 reads that had at least a 90% match to the preliminary barcode. These reads would then be used to call another barcode with MAFFT. Only if this also failed to produce a QC-compliant barcode, ONTbarcoder would “fix” the preliminary barcode using its 20 closest matches in the dataset. 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).* 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).

After obtaining the barcodes, we first investigated barcode accuracy (Figure 1) by directly aligning the MinION barcodes with the corresponding Sanger and Illumina barcodes. We find that MinION barcodes are virtually identical to Sanger and Illumina barcodes (>99.99% identity, Table 3). We then established that the number of ambiguous bases ("N") is also 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 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 easily match the Consortium for the Barcode of Life (CBOL) criteria for "barcode" designation with regard to length, accuracy, and ambiguity.

Rarefaction at the different coverages reveals that 80-90% of high-quality barcodes are obtained within a few hours of sequencing and that the number of barcodes generated by MinION was higher or comparable to what could be obtained with Sanger or Illumina sequencing (Figure 1). We can use the same data to determine the coverage needed for obtaining reliable barcodes. For this purpose, we plotted the results using coverage on the x-axis instead of time (Figure 2). This reveals that the vast majority of specimen bins yield high-quality barcodes at coverages between 25x and 50x when R10.3 reads are used. Increasing coverage beyond 50x leads to only modest improvements of barcode quality and few additional specimen amplicons yield new barcodes. The coverage needed for obtaining Flongle barcodes is somewhat higher, but the main difference between the R9.4 technology of the Flongle flow cell and R10.3 is that more barcodes retain ambiguous bases even at high coverage. The differences in read quality between R9.4 and R10.3 become even more obvious when the read bins for the "Mixed Diptera subsample" are 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, this means that 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 contaminants being called. For example, we observed that some “negative” PCR controls were starting to yield low-quality barcodes for 4 of 105 negatives in the Palaearctic Phoridae (313 bp) and 1 of 104 negatives in the Palaearctic Phoridae (658 bp) datasets.

To facilitate the planning of barcode projects, we illustrate the trade-offs between barcode yield, time, and amount of raw data needed for six amplicon pools (Figure 4: 191-9,934 specimens). These standard curves can be used to roughly estimate the amount of data 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). For each dataset, we illustrate how much data were needed to recover a certain proportion of barcodes. The number of recoverable barcodes was set to the number of all error-free, filtered barcodes (category 2) 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 much data are needed to meet barcoding targets because, for example, the demultiplexing rates differ between flow cells and different amplicon pools have very different read abundance distributions (see Figure 3).

Table 3. Quality assessment of barcodes generated by ONTbarcoder at 200X 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. All denominators for calculating percentages 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 are likely to represent contaminations and would not represent per base accuracy. This procedure was followed also in Srivathsan, Hartop et al. (2019).

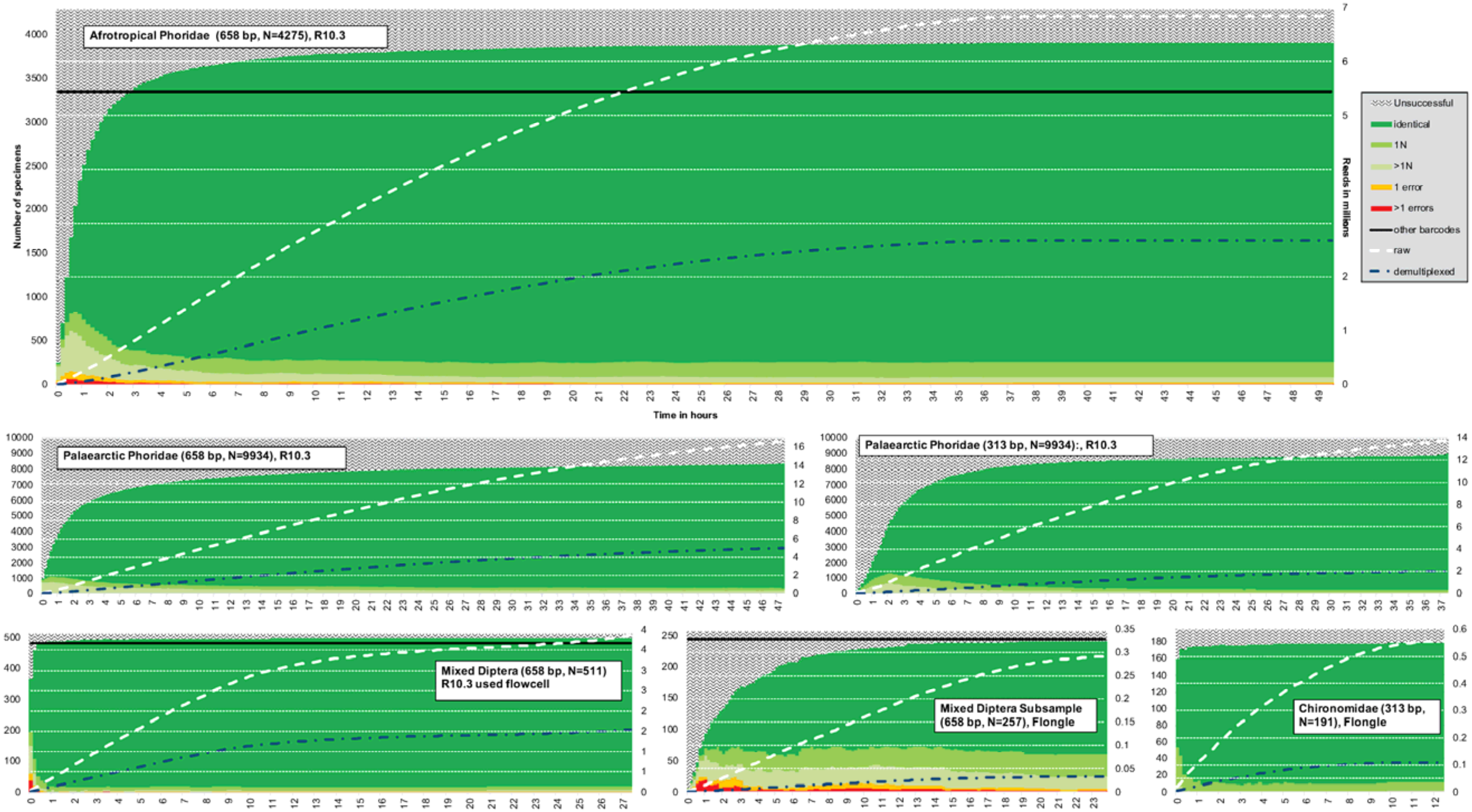

Figure 1. Rapid recovery of accurate MinION barcodes over time (in hours, x-axis) (filtered barcodes: dark green = barcodes passing all 4 QC criteria, light green = one ambiguous base; lighter green = more than 1N, no barcode = white with pattern, 1 mismatch = orange, >1 mismatch = red). The solid black 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)

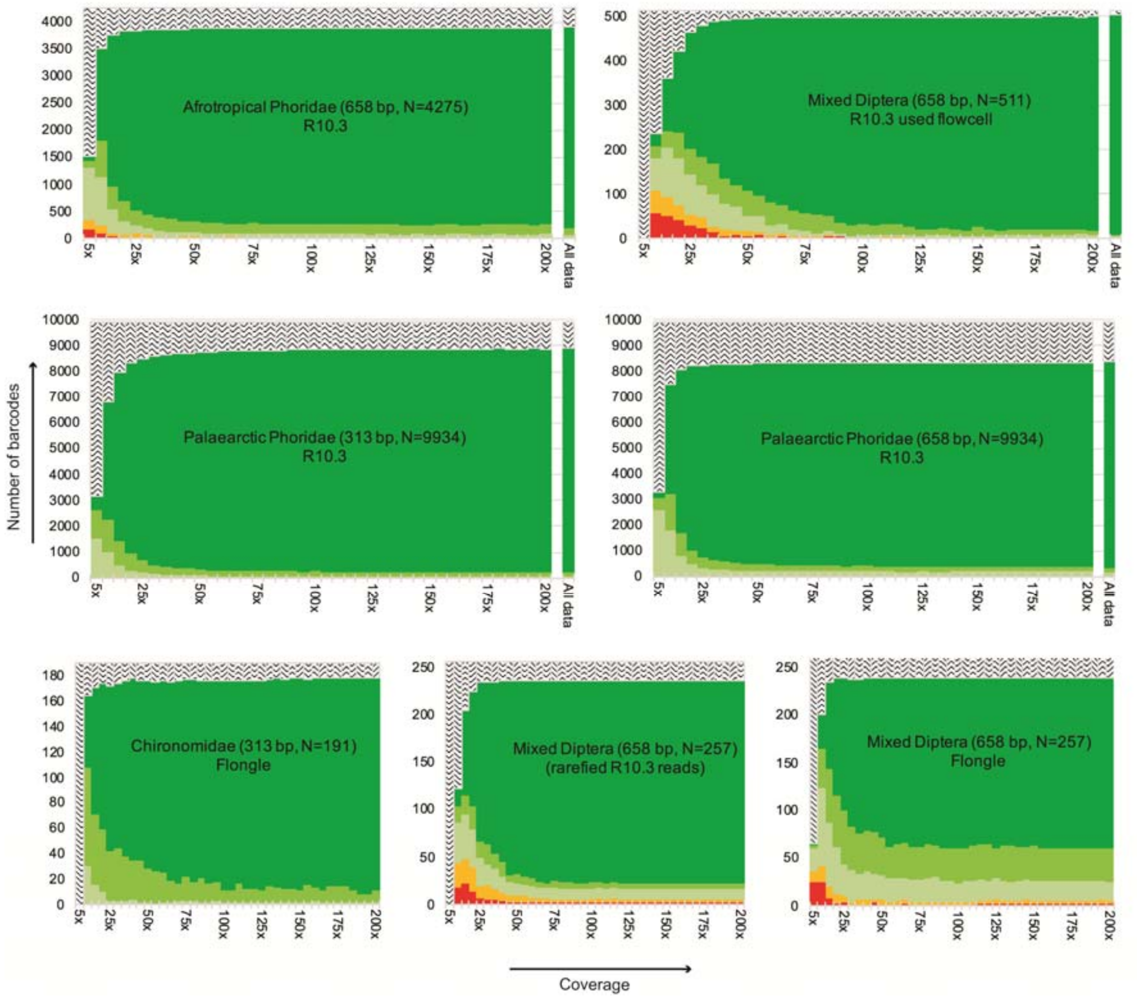

Figure 2. Relationship between barcode quality and coverage. Subsetting the data to 5-200X coverage shows that there are very minor gains to barcode quality after 25-50X coverage. (filtered barcodes: dark green = barcodes passing all 4 QC criteria, light green = one ambiguous base; lighter green = more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

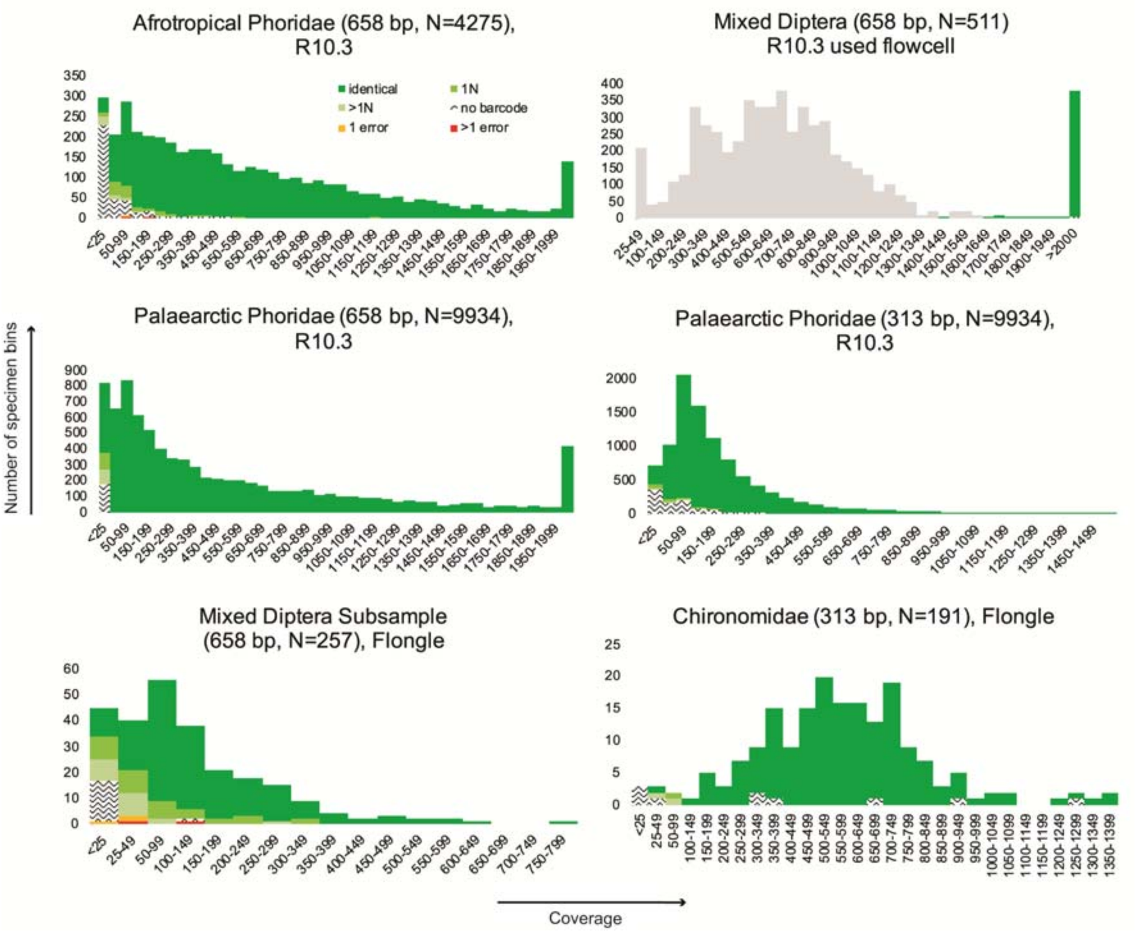

Figure 3. Bin size distribution for six amplicon pools (color-coding as in Figs 1-2). Due to overly generous coverage for the “Mixed Diptera” dataset, we use grey to show the bin size distribution after dividing the bin read totals by 5.

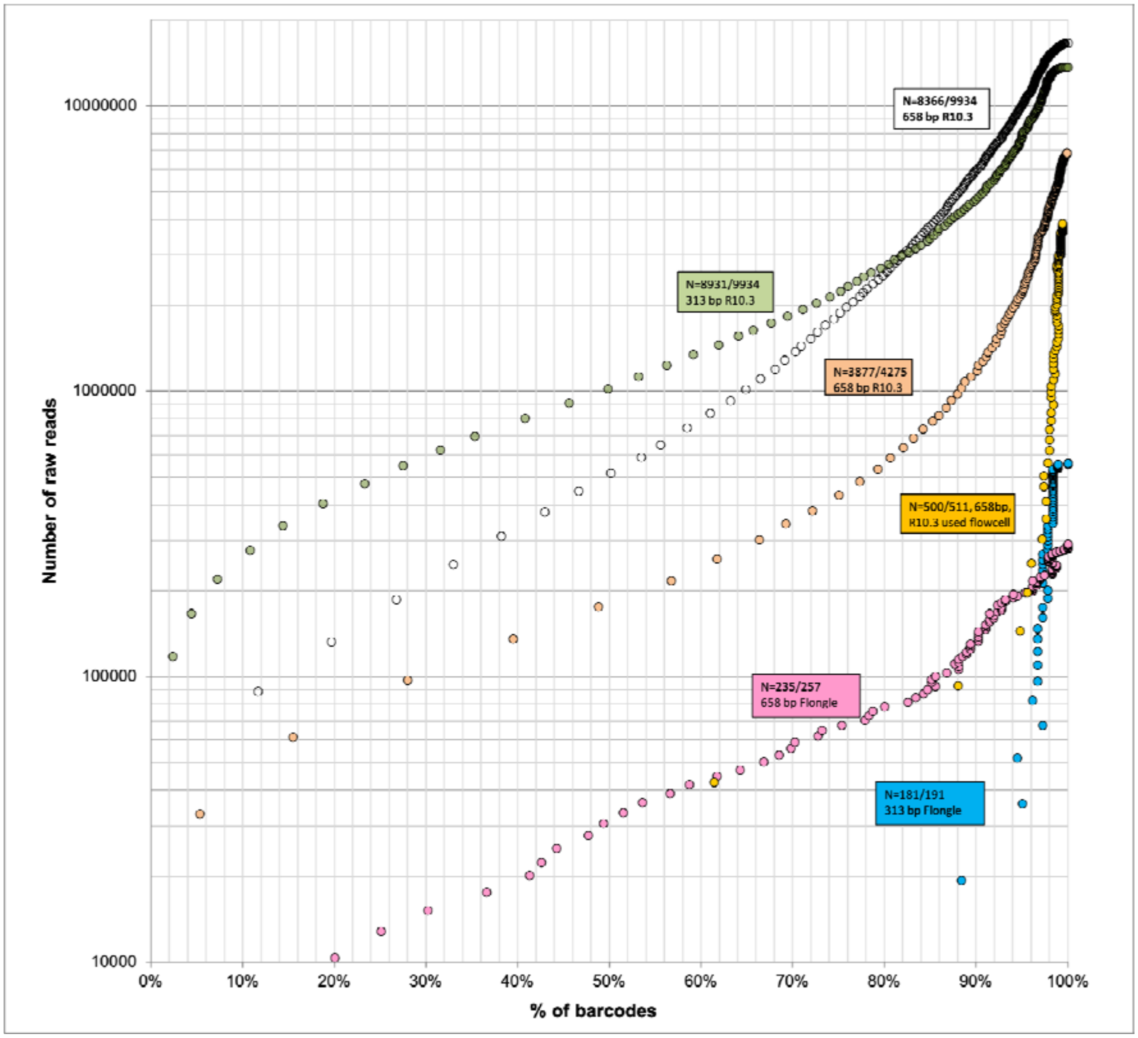

Figure 4. Relationship between barcoding success and number of raw reads for six amplicon pools (191-9934 specimens; barcoding success rates 84-97%). Percentage of barcodes recovered is relative to the final estimate based on all data.

Discussion

Biodiversity research needs new scalable techniques for large-scale species. 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. Generating DNA barcodes involves three processes. The first is obtaining a DNA template, and we have herein outlined some simplified procedures that render this process essentially free-of-cost, although automation and AI-based solutions will be useful for processing very large numbers of specimens in countries with high manpower cost. The second step is getting tagged amplicons via PCR. We here also described simplified procedures, but further simplification is possible. For example, the use of hydrocyclers and/or 384-well plates can further reduce PCR costs. Traditionally, this second step in the barcoding process has been somewhat neglected because the main obstacle to cost-effective barcoding was the third step; i.e., the sequencing of the amplicon. Fortunately, there are now several cost-effective solutions based on 2nd and 3rd generation sequencing technologies.

We here argue that sequencing with MinION is particularly attractive. Library preparation can be learned within hours and an automated library preparation instrument is in development that will eventually work for ligation-based libraries. Furthermore, MinION flow cells can accommodate projects of greatly 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 enough have been acquired. Flow cells can then be washed and re-used again. However, with each use the remaining capacity of the flow cell declines because some nanopores will become unavailable. Eventually, too few pores remain active and the flow cell will be spent. Traditionally, the main obstacles to using MinION have been poor read quality and high cost. Fortunately, both issues seem to be 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 no longer a need to polish reads or rely on external data or algorithms. The greater ease with which MinION barcodes can be obtained are due to several factors. Firstly, much larger numbers of reads can now be obtained with one MinION flow cell. 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 yield very accurate barcodes; i.e., both barcode quality and quantity are greatly improved.

We previously tested MinION 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 used the remaining ~500 pores of a used R10.3 flow cell that was run for 49 hours when used for the first time. After washing, we obtained 502 barcodes and >98% (496) of them were 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 reconstruct 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 (Srivathsan, Hartop et al. 2019). Now, using 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, MinION will soon be the default barcoding tool for many 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 flow cell (Ho, Puniamoorthy et al. 2020). This democratizes biodiversity discovery and allows many biologists, government agencies, students, and citizen scientists from around the globe to get involved in these initiatives. Biodiversity discovery with cost-effective barcodes will also facilitate biodiversity discovery in countries with high biodiversity but limited science funding.

Table 4. Equipment required for MinION barcoding

Required	
1	Thermocycler(s)
2	Gel Electrophoresis setup
3	Magnetic Separation Rack
4	Vortex
5	Mini-centrifuge
6	MinION sequencer
7	Freezer and fridge
8	Qubit for DNA quantification
Optional but highly desirable	
1	Multichannel pipette(s)
2	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 target a 80% success rate in order to identify the dominant species in a sample. Based on Figure 4, only ca. 1.5 million raw MinION reads would be needed. Given that 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. Including a library cost of ca. USD 100 (kit includes chemistry for six libraries), the overall sequencing cost of a project that requires 1.5 million reads 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. However, this will then 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., additional sequencing yields diminishing returns. Better gains will be made if failed barcodes are re-pooled and re-sequenced as done by Srivathsan, Hartop et al. (2019). Overall, we predict that most users will, at most, try to multiplex 10,000 amplicons in the same MinION flow cell. However, we also predict that large-scale biodiversity projects will switch to sequencing with PromethION, a larger sequencing unit that can accommodate up to 48 flow cells. This will lower the sequencing cost by more than 60%, as PromethION flow cells have 6 times the number of pores for twice the cost (capacity per flow cell should be 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, but Flongle barcodes will remain comparatively expensive because each flow cell costs $90 and requires a library that is prepared with half the normal reagents (ca. $50). A change of the flow cell chemistry from that of R9.4 to R10.3 would, however, help with improving the quality of the barcodes obtained from Flongle. Lastly the initial setup cost for MinION/Flongle, can be as low as 1000 USD, but we recommend purchase of Mk1C unit at 4900 USD for easy access to GPU required for high accuracy basecalling. Obtaining flow cells at low cost often requires collaboration between several labs because it allows for buying flow cells in bulk.

There are a number of studies that have used MinION for barcoding fungi, animals, and plants (Menegon, Cantaloni et al. 2017, Pomerantz, Peñafiel et al. 2018, Wurzbacher, Larsson et al. 2018, Krehenwinkel, Pomerantz et al. 2019, Maestri, Cosetino et al. 2019, Chang, Ip et al. 2020, Chang, Ip et al. 2020, Knot, Zouganelis et al. 2020, Seah, Lim et al. 2020, Sahlin, Lim et al. 2021). There is one fundamental difference between these studies and the vision presented here. These studies tended to focus on the use of MinION sequencing in the field and only a very small number of specimens were analysed (<150 with the exception of >500 in Chang, Ip et al (2020)). The use of MinION in the field is an attractive feature of the technology, especially for time-sensitive samples that could degrade before reaching a lab. However, it is unlikely to help substantially with tackling the challenges related to large-scale biodiversity discovery and monitoring. Small-scale projects carried out in the field with MinION yield barcodes that are so expensive they are too expensive for most researchers in biodiverse countries. Additionally, the bioinformatic pipelines that were developed for these small-scale projects were not suitable for large-scale, decentralized barcoding in a large variety of facilities. For example, some of the studies used ONT's commercial barcoding kit that only allows for multiplexing up to 96 samples in one flow cell (Maestri, Cosetino et al. 2019, Seah, Lim et al. 2020); i.e., each amplicon had very high read coverage which influenced the corresponding bioinformatics pipelines (e.g. ONTrack's recommendation is 1000x: (Maestri, Cosentino et al. 2019). The generation of such high coverage datasets also meant that the pipelines were only tested for so few samples (<60: (Menegon, Cantaloni et al. 2017, Maestri, Cosetino et al. 2019, Seah, Lim et al. 2020, Sahlin, Lim et al. 2021) that these tests were unlikely to represent the complexities of large, multiplexed amplicon pools (e.g., nucleotide diversity, uneven coverage).

ONTbarcoder evolved from miniBarcoder, which was utilized in four studies covering >7000 barcodes (Srivathsan, Baloglu et al. 2018, Srivathsan, Hartop et al. 2019, Chang, Ip et al. 2020, Chang, Ip et al. 2020). The new software introduced here addresses two drawbacks of its precursor, miniBarcoder. (1) The latter used a 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. Fortunately, such errors are now exceedingly rare. (2) miniBarcoder also had several external dependencies including RACON, GraphMap, BLAST, glsearch36 (Sović, Šikić et al. 2016, Pearson 2017, Vaser, Sovic et al. 2017) which made installation difficult and limited its usage on computers running Windows. Such dependencies on external software are a drawback of all MinION bioinformatics pipelines prior to ONTbarcoder. For example, the one described by (Sahlin, Lim et al. 2021) involves minibar/qcat and nanofilt, while NGSpeciesID relies on isONclust SPOA, Parasail, and optionally, Medaka (Daily 2016, Krehenwinkel, Pomerantz et al. 2019, Sahlin and Medvedev 2020). These dependencies and complexities meant that Watsa et al. (2020) recommended bioinformatics training before MinION barcoding could be used in schools (e.g., training in UNIX command-line) and additionally required the installation of several software tools onto the teaching computers. Neither is needed for ONTbarcoder, which runs on a regular laptop and has been extensively tested (>4000 direct comparisons to Sanger and Illumina barcodes). In addition, ONTbarcoder is designed in a way that thousands of barcodes can be obtained rapidly without impairing accuracy; i.e., one can run a very fast analysis by using low read coverage, but fewer barcodes would be recovered because many would not pass the 4 QC criteria. Speed is also achieved through the parallelization of most steps on UNIX systems (Mac and Linux) (parallelization is restricted to demultiplexing in Windows). Based on the recent past, we expect many MinION to continue to evolve quickly. We expect flow cell capacity to increase further and basecalling to improve (see (Xu, Mai et al. 2020). Currently, the main limitation for MinION barcoding is still the slow speed of high accuracy basecalling on the MinION MK1C, the ONT instrument most suitable for the average user.

Some readers are likely to argue that large-scale biodiversity discovery and monitoring can be more efficiently carried out via metabarcoding of whole samples consisting of hundreds or thousands of specimens. This would question the need for large-scale, decentralized barcoding of individual specimens. However, large-scale barcoding and metabarcoding will more likely complement each other. For example, large-scale barcoding of individual specimens remains essential for discovering and describing species. It is important to remember that *COI* lumps recently diverged species and divides species with deep allopatric splits (Hickerson, Meyer et al. 2006), making the ability to relate barcodes to individual specimens critical for barcode cluster validation. The reasons for these complications are well understood and include introgression, lineage sorting, and long periods of allopatry within species. It is therefore not advisable to identify or describe species based on *COI* sequences only. Ignoring these shortcomings of DNA barcodes will also negatively impact the likelihood of obtaining accurate species-level resolution from the analysis of metabarcoding data. Such data is best analyzed using comprehensive barcode databases that contain species-level information and *COI* sequences from different clades. High quality barcode databases are important for the analysis of metabarcoding data because they facilitate the identification of numts, heteroplasmy, contaminants and errors. Large-scale barcoding will also be needed in order to benefit from another new technique that may become critical for biodiversity discovery and monitoring; i.e. AI-assisted analysis of images (Valan, Makonyi et al. 2019). Large-scale barcoding generates identified specimens that can be imaged and utilized for training neural networks. With increasing advancements in imaging hardware, computational processing power and machine learning systems, AI-assisted biodiversity monitoring could be the method of choice in the future because it could quickly determine and count many common species and only specimens from new/rare species would still require barcoding.

Conclusions

Many biologists would like to have ready access to barcodes without having to run large and complex laboratories or send specimens halfway around the world. Many have also been impressed by MinION's low cost, portability, and ability to deliver real-time sequencing, but large-scale barcoding with MinION has yet to get established due to previously high costs 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 for researchers with little bioinformatics background to analyze the data on a standard laptop. Our simplified techniques for obtaining barcode amplicons save time and research funding, and makes biodiversity discovery scalable and accessible to all.

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. This work was supported by a Ministry of Education grant on biodiversity discovery (R-154-000-A22-112).

Software and test dataset availability

ONTbarcoder is available at https://github.com/asrivathsan/ONTbarcoder, which also contains the link to download the test files.

Literature cited

Ahrens, D., T. Fujisawa, H. J. Krammer, J. Eberle, S. Fabrizi and A. P. Vogler (2016). "Rarity and incomplete sampling in DNA-based species delimitation " Systematic Biology **65**(3): 478-494.

Arribas, P., C. Andújar, K. Hopkins, M. Shepherd and A. P. Vogler (2016). "Metabarcoding and mitochondrial metagenomics of endogean arthropods to unveil the mesofauna of the soil." Methods in Ecology and Evolution **7**(9): 1071-1081.

Baloğlu, B., E. Clews and R. Meier (2018). "NGS barcoding reveals high resistance of a hyperdiverse chironomid (Diptera) swamp fauna against invasion from adjacent freshwater reservoirs." Frontiers in Zoology **15**(1): 31.

Bar-On, Y. M., R. Phillips and R. Milo (2018). "The biomass distribution on Earth." Proceedings of the National Academy of Sciences **115**(25): 6506-6511.

Barrett, R. D. H. and P. D. Hebert (2005). "Identifying spiders through DNA barcodes." Canadian Journal of Zoology **83**: 481-491.

Chang, J. J. M., Y. C. A. Ip, A. G. Bauman and D. Huang (2020). "MinION-in-ARMS: Nanopore sequencing to expedite barcoding of specimen-rich macrofaunal samples from autonomous reef monitoring structures." Frontiers in Marine Science **7**: 448.

Chang, J. J. M., Y. C. A. Ip, C. S. L. Ng and D. Huang (2020). "Takeaways from mobile DNA barcoding with BentoLab and MinION." Genes **11**: 1121.

Crampton-Platt, A., D. W. Yu, X. Zhou and A. P. Vogler (2016). "Mitochondrial metagenomics: letting the genes out of the bottle." Gigascience **5**(1): s13742-13016-10120-y.

Daily, J. (2016). "Parasail: SIMD C library for global, semi-global, and local pairwise sequence alignments." BMC Bioinformatics **17**: 81.

Forum, W. E. (2020). "World Economic Forum. The Global Risks Report 2020.", from https://www.weforum.org/reports/the-global-risks-report-2020.

Frith, M. C., S. Mitsuhashi and K. Katoh (2020). lamassemble: Multiple Alignment and Consensus Sequence of Long Reads. Multiple Sequence Alignment. K. Katoh. New York, Humana**:** 135-145.

Groombridge, B., Ed. (1992). Global Biodiversity: Status of the Earth's Living Resources. World Conservation Monitoring Centre. London, Chapman & Hall.

Grootaert, P. (2018). "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 **66**: 413-473.

Grootaert, P. (2019). "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 **554**: 1-27.

Hamada, M., Y. Ono, K. Asai and M. C. J. B. Frith (2017). "Training alignment parameters for arbitrary sequencers with LAST-TRAIN." **33**(6): 926-928.

Hebert, P. D., T. W. A. Braukmann, S. W. J. Prosser, S. Ratnasingham, J. R. deWaard, N. V. Ivanova, D. Janzen, W. Hallwachs, S. Naik, J. E. Sones and E. V. Zakharov (2018). "A Sequel to Sanger: amplicon sequencing that scales." BMC Genomics **19**: 219.

Hebert, P. D., J. R. DeWaard, E. V. Zakharov, S. W. J. Prosser, J. E. Sones, J. T. A. McKeown, B. Mantle and J. La Salle (2013). "A DNA 'Barcode Blitz': Rapid digitization and sequencing of a Natural History collection." PLoS One **8**(7): e68535.

Hebert, P. D. N., A. Cywinska, S. L. Ball and J. R. deWaard (2003). "Biological identifications through DNA barcodes." Proceedings of the Royal Society Biological Sciences Series B **270**(1512): 313-321.

Hebert, P. D. N., S. Ratnasingham, E. V. Zakharov, A. C. Telfer, V. Levesque-Beaudin, M. A. Milton, S. Pedersen, P. Jannetta and J. R. deWaard (2016). "Counting animal species with DNA barcodes: Canadian insects." Philosophical Transactions of the Royal Society B: Biological Sciences **371**: 20150333.

Hendrich, L., J. Pons, I. Ribera and M. Balke (2010). "Mitochondrial Cox1 sequence data reliably uncover patterns of insect diversity but suffer from high lineage-idiosyncratic error rates." PLoS One **5**(12): e14448.

Hickerson, M. J., C. P. Meyer and Moritz (2006). "DNA barcoding will often fail to discover new animal species over broad parameter space." Systematic Biology **55**(5): 729-739.

Ho, J. K. I., J. Puniamoorthy, A. Srivathsan and R. Meier (2020). “MinION sequencing of seafood in Singapore reveals creatively labelled flatfishes, confused roe, pig DNA in squid balls, and phantom crustaceans.” Food Control **112:** 107144.

Ismay, B. and Y. Ang (2019). "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 **67**: 412-420.

Ivanova, N. V., A. V. Borisenko and P. D. N. Hebert (2009). "Express barcodes: racing from specimen to identification." Molecular Ecology Resources **9**: 35-41.

Ivanova, N. V., J. R. Dewaard and P. D. N. Hebert (2006). "An inexpensive, automation-friendly protocol for recovering high-quality DNA." Molecular Ecology Notes **6**(4): 998-1002.

Katoh, K. and D. M. Standley (2013). "MAFFT Multiple Sequence Alignment Software Version 7: Improvements in performance and usability." Molecular Biology and Evolution **30**(4): 772-780.

Kekkonen, M., M. Mutanen, L. Kaila, M. Nieminen and P. D. Hebert (2015). "Delineating Species with DNA Barcodes: A case of taxon dependent method performance in moths." PLoS One **10**(4): e0122481.

Knot, I. E., G. D. Zouganelis, G. D. Weedall, S. A. Wich and R. Rae (2020). "DNA barcoding of nematodes using the MinION." Frontiers in Ecology and Evolution **8**: 100.

Knox, M. A., I. D. Hogg, C. A. Pilditch, J. C. Garcia-R, P. D. N. Hebert and D. Steinke (2020). "Contrasting patterns of genetic differentiation for deep-sea amphipod taxa along New Zealand's continental margins." Deep Sea Research Part I: Oceanographic Research Papers **162**: 103323.

Kranzfelder, P., T. Ekrem and E. Stur (2016). "Trace DNA from insect skins: a comparison of five extraction protocols and direct PCR on chironomid pupal exuviae." Molecular Ecology Resources **16**(1): 353-363.

Krehenwinkel, H., S. R. Kennedy, A. Rueda, A. Lam and R. G. Gillespie (2018). "Scaling up DNA barcoding – Primer sets for simple and cost efficient arthropod systematics by multiplex PCR and Illumina amplicon sequencing." Methods in Ecology and Evolution **9**(11): 2181-2193.

Krehenwinkel, H., A. Pomerantz, J. B. Henderson, S. R. Kennedy, J. Y. Lim, V. Swamy, J. D. Shoobridge, N. Graham, N. H. Patel, R. G. Gillespie and S. Prost (2019). "Nanopore sequencing of long ribosomal DNA amplicons enables portable and simple biodiversity assessments with high phylogenetic resolution across broad taxonomic scale." Gigascience **8**(5): giz006.

Krell, F. T. (2004). "Parataxonomy vs. taxonomy in biodiversity studies - pitfalls and applicability of 'morphospecies' sorting." Biodiversity and Conservation **13**(4): 795-812.

Kwong, S., A. Srivathsan and R. Meier (2012). "An update on DNA barcoding: low species coverage and numerous unidentified sequences." Cladistics **28**(6): 639-644.

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

Maestri, S., E. Cosetino, M. Paterno, H. Freitag, J. M. Garces, L. Marcolungo, M. Alfano, I. Njunjić, M. Schilthuizen, F. Slik, M. Menegon, M. Rossato and M. Delledonne (2019). "A rapid and accurate MinION-based workflow for tracking species biodiversity in the field." Genes **10**(6): 468.

Meier, R. (2008). DNA sequences in taxonomy - Opportunities and challenges. New Taxonomy. Q. D. Wheeler. **76:** 95-127.

Meier, R., W. H. Wong, A. Srivathsan and M. S. Foo (2016). "$1 DNA barcodes for reconstructing complex phenomes and finding rare species in specimen-rich samples." Cladistics **32**(1): 100-110.

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

Ng'endo, R. N., Z. B. Osiemo and R. Brandl (2013). "DNA barcodes for species identification in the hyperdiverse ant genus *Pheidole* (Formicidae: Myrmicinae)." Journal of Insect Science **13**: 27.

Page, R. (2011). "Dark taxa: GenBank in a post-taxonomic world." https://iphylo.blogspot.com/2011/04/dark-taxa-genbank-in-post-taxonomic.html, Accessed February 2021.

Pearson, W. R. (2017). "Finding protein and nucleotide similarities with FASTA." Current Protocols in Bioinformatics **53**: 3.9.1-3.9.25.

Pomerantz, A., N. Peñafiel, A. Arteaga, L. Bustamante, F. Pichardo, L. A. Coloma, C. L. Barrio-Amorós, D. Salazar-Valenzuela and S. J. G. Prost (2018). "Real-time DNA barcoding in a rainforest using nanopore sequencing: opportunities for rapid biodiversity assessments and local capacity building." **7**(4): giy033.

Ponder, W. and D. Lunney (1999). The Other 99% - the Conservation and Biodiversity of Invertebrates. Sydney, Transactions of the Royal Zoological Society of New South Wales.

Swiss Re. (2020). " Biodiversity and Ecosystem Services A business case for re/insurance." Zurich, Swiss Re Management Ltd.

Rohland, N. and D. Reich (2012). "Cost-effective, high-throughput DNA sequencing libraries for multiplexed target capture." Genome research **22**: 939-946.

Sahlin, K., M. C. W. Lim and S. Prost (2021). "NGSpeciesID: DNA barcode and amplicon consensus generation from long-read sequencing data." Ecology and Evolution **11**(3): 1392-1398.

Sahlin, K. and P. Medvedev (2020). "De novo clustering of long-read transcriptome data using a greedy, quality value-based algorithm." Journal of Computational Biology **27**(4): 472-484.

Samoh, A., C. Satasook and P. Grootaert (2019). "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 **67**: 640-659.

Seah, A., M. C. W. Lim, D. McAloose, S. Prost and T. A. Seimon (2020). "MinION-based DNA barcoding of preserved and non-Invasively vollected wildlife samples." Genes **11**(4): 445.

Shokralla, S., T. M. Porter, J. F. Gibson, R. Dobosz, D. Janzen, W. Hallwachs, G. B. Golding and M. Hajibabaei (2015). "Massively parallel multiplex DNA sequencing for specimen identification using an Illumina MiSeq platform." Scientific Reports **5**: 9687.

Shokralla, S., J. L. Spall, J. F. Gibson and M. Hajibabaei (2012). "Next-generation sequencing technologies for environmental DNA research." Molecular Ecology **21**(8): 1794-1805.

Silvestre-Ryan, J. and I. Holmes (2021). "Pair consensus decoding improves accuracy of neural network basecallers for nanopore sequencing." Genome Biology **22**: 38.

Sović, I., M. Šikić, A. Wilm, S. N. Fenlon, S. Chen and N. Nagarajan (2016). "Fast and sensitive mapping of nanopore sequencing reads with GraphMap." Nature Communications **7**: 11307.

Srivathsan, A., B. Baloğlu, W. Wang, W. X. Tan, D. Bertrand, A. H. Q. Ng, E. J. H. Boey, J. J. Y. Koh, N. Nagarajan and R. Meier (2018). "A MinION-based pipeline for fast and cost-effective DNA barcoding." Molecular Ecology Resources **18**(5): 1035-1049.

Srivathsan, A., E. Hartop, J. Puniamoorthy, W. T. Lee, S. N. Kutty, O. Kurina and R. Meier (2019). "Rapid, large-scale species discovery in hyperdiverse taxa using 1D MinION sequencing." BMC Biology **17**(1): 96.

Srivathsan, A., N. Nagarajan and R. Meier (2019). "Boosting natural history research via metagenomic clean-up of crowdsourced feces." PLoS Biology **17**(11): e3000517.

Stork, N. E., J. McBroom, C. Gely and A. J. Hamilton (2015). "New approaches narrow global species estimates for beetles, insects, and terrestrial arthropods." Proceedings of the National Academy of Sciences **112**(24): 7519-7523.

Stribling, J. B., K. L. Pavlik, S. M. Holdsworth and E. W. Leppo (2008). "Data quality, performance, and uncertainty in taxonomic identification for biological assessments." Journal of the North American Benthological Society **27**(4): 906-919.

Tang, C. F., P. Grootaert and D. Yang (2018). "*Protomedetera*, a new genus from the Oriental and Australasian realms (Diptera, Dolichopodidae, Medeterinae)." Zookeys **743**: 137-151.

Tang, C. F., D. Yang and P. Grootaert (2018). "Revision of the genus *Lichtwardtia* Enderlein in Southeast Asia, a tale of highly diverse male terminalia (Diptera, Dolichopodidae)." Zookeys **798**: 63-107.

Tautz, D., P. Arctander, A. Minelli, R. H. Thomas and A. P. Vogler (2003). "A plea for DNA taxonomy." Trends in Ecology & Evolution **18**(2): 70-74.

Thongjued, K., W. Chotigeat, S. Bumrungsri, P. Thanakiatkrai and T. Kitpipit (2019). "A new cost-effective and fast direct PCR protocol for insects based on PBS buffer." Molecular Ecology Resources **19**(3): 691-701.

Thormann, B., D. Ahrens, D. M. Armijos, M. K. Peters and T. Wagner (2016). "Exploring the leaf beetle fauna (Coleoptera: Chrysomelidae) of an Ecuadorian mountain forest using DNA barcoding." PLoS One **11**(2): e0148268.

Truett, G., P. Heeger, R. Mynatt, A. Truett, J. Walker and M. J. B. Warman (2000). "Preparation of PCR-quality mouse genomic DNA with hot sodium hydroxide and tris (HotSHOT)." Biotechniques **29**(1): 52-54.

Valan, M., K. Makonyi, A. Maki, D. Vondráček and F. Ronquist (2019). "Automated taxonomic identification of insects with expert-level accuracy using effective feature transfer from convolutional networks." Systematic Biology **68**(6): 876-895.

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

Vereecke, N., J. Bokma, F. Haesebrouck, H. Nauwynck, F. Boyen, B. Pardon and S. Theuns (2020). "High quality genome assemblies of *Mycoplasma bovis* using a taxon-specific Bonito basecaller for MinION and Flongle long-read nanopore sequencing." BMC Bioinformatics **21**: 517.

Wang, W. Y., A. Srivathsan, M. Foo, S. K. Yamane and R. Meier (2018). "Sorting specimen-rich invertebrate samples with cost-effective NGS barcodes: Validating a reverse workflow for specimen processing." Molecular Ecology Resources **18**(3): 490-501.

Wang, W. Y., A. Yamada and K. Eguchi (2018). "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 **66**: 652-663.

Wang, W. Y., A. Yamada and S. Yamane (2020). "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 **915**: 137-174.

Wang, W. Y., G. W. J. Yong and W. Jaitrong (2018). "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 **4532**(3): 301-340.

Watsa, M., G. A. Erkenswick, a. Pomerantz and S. Prost (2020). "Portable sequencing as a teaching tool in conservation and biodiversity research." PLoS Biology **18**(4): e3000667.

Wick, R. R. (2019). "Performance of neural network basecalling tools for Oxford Nanopore sequencing." Genome Biology **20**: 129.

Wong, W. H., Y. C. Tay, J. Puniamoorthy, M. Balke, P. S. Cranston and R. Meier (2014). "'Direct PCR' optimization yields a rapid, cost-effective, nondestructive and efficient method for obtaining DNA barcodes without DNA extraction." Molecular Ecology Resources **14**(6): 1271-1280.

Wurzbacher, C., E. Larsson, J. Bengtsson-Palme, S. V. den Wyngaert, S. Svantesson, E. Kristiansson, M. Kagami and R. H. Nilsson (2018). " Introducing ribosomal tandem repeat barcoding for fungi." Molecular Ecology Resources **19**(1): 118-127.

Xu, Z., Y. Mai, D. Liu, W. He, X. Lin, C. Xu, L. Zhang, X. Meng, J. Mafofo, W. A. Zaher, Y. Li and N. Qiao (2020). "Fast-Bonito: A faster basecaller for nanopore sequencing." BioRxiv: doi:10.1101/2020.1110.1108.318535.

Yeo, D., J. Puniamoorthy, R. W. J. Ngiam and R. Meier (2018). "Towards holomorphology in entomology: rapid and cost-effective adult-larva matching using NGS barcodes." Systematic Entomology **43**(4): 678-691.

Yeo, D., A. Srivathsan and R. Meier (2020). "Longer is Not Always Better: Optimizing Barcode Length for Large-Scale Species Discovery and Identification." Systematic Biology **69**(5): 999-1015.

Yeo, D., A. Srivathsan, J. Puniamoorthy, M. Foo, P. Grootaert, L. Chan, B. Guenard, C. Damken, R. A. Wahab and Y. J. b. Ang (2020). "Mangroves are an overlooked hotspot of insect diversity despite low plant diversity." BioRxiv: doi:10.1101/2020.12.17.423191.