Yang et al. BMC Genomics (2020) 21:862
https://doi.org/10.1186/s12864-020-07255-w

BMC Genomics

METHODOLOGY ARTICLE Open Access

Efficient COI barcoding using high throughput single-end 400 bp sequencing

Chentao Yang[1], Yuxuan Zheng[2], Shangjin Tan[1], Guanliang Meng[1], Wei Rao[1], Caiqing Yang[2], David G. Bourne[3,4,5], Paul A. O'Brien[3,4,5], Junqiang Xu[1], Sha Liao[1], Ao Chen[1], Xiaowei Chen[1], Xinrui Jia[2], Ai-bing Zhang[2*] and Shanlin Liu[1,6*]

Abstract

Background: Over the last decade, the rapid development of high-throughput sequencing platforms has accelerated species description and assisted morphological classification through DNA barcoding. However, the current high-throughput DNA barcoding methods cannot obtain full-length barcode sequences due to read length limitations (e.g. a maximum read length of 300 bp for the Illumina's MiSeq system), or are hindered by a relatively high cost or low sequencing output (e.g. a maximum number of eight million reads per cell for the PacBio's SEQUEL II system).

Results: Pooled cytochrome c oxidase subunit I (*COI*) barcodes from individual specimens were sequenced on the MGISEQ-2000 platform using the single-end 400 bp (SE400) module. We present a bioinformatic pipeline, HIFI-SE, that takes reads generated from the 5′ and 3′ ends of the *COI* barcode region and assembles them into full-length barcodes. HIFI-SE is written in Python and includes four function modules of *filter*, *assign*, *assembly* and *taxonomy*. We applied the HIFI-SE to a set of 845 samples (30 marine invertebrates, 815 insects) and delivered a total of 747 fully assembled *COI* barcodes as well as 70 *Wolbachia* and fungi symbionts. Compared to their corresponding Sanger sequences (72 sequences available), nearly all samples (71/72) were correctly and accurately assembled, including 46 samples that had a similarity score of 100% and 25 of ca. 99%.

Conclusions: The HIFI-SE pipeline represents an efficient way to produce standard full-length barcodes, while the reasonable cost and high sensitivity of our method can contribute considerably more DNA barcodes under the same budget. Our method thereby advances DNA-based species identification from diverse ecosystems and increases the number of relevant applications.

Keywords: DNA barcode, High-throughput sequencing, MGISEQ-2000, SE400, *COI*, Biodiversity

Background

Since it was first proposed by Hebert et al. [1], DNA barcoding has attracted global synergistic efforts resulting in well-curated and centralized reference databases. The Barcode of Life Data systems (BOLD) [2], for example, has been growing into a repository of greater than 11 M barcodes representing 314 K species (accessed in Jun. 2020). The applications of DNA barcoding are wide-ranging and may be used to identify species across different life stages and from various environments (e.g. predator feces [3, 4] and from stomach contents [5]). This, along with the ease of barcoding accessibility and analysis, has led to its use in a wide spectrum of scientific and commercial areas, such as cryptic species discovery [6], biodiversity monitoring [7–9], conservation biology [10], inspection of illegal trade of endangered species [11] and discovery of illegal ingredients in medicine [12].

* Correspondence: zhangab2008@mail.cnu.edu.cn; shanlin1115@gmail.com
[2]College of Life Sciences, Capital Normal University, Beijing 100048, China
[1]BGI-Shenzhen, Shenzhen 518083, China
Full list of author information is available at the end of the article

Barcode sequences have been accumulating rapidly in the last decade, prompting a need to improve the available reference databases as they are currently limited by poor and biased spatial coverage and skewed taxonomic coverage [13–16]. Biodiversity initiatives are often limited by insufficient funding, which makes it difficult to include both morphological identification and DNA-based taxonomic work. Therefore, scientists have been attempting to generate cost-efficient barcode sequences via high-throughput sequencing (HTS) platforms. Reduced costs would increase the accessibility of large-scale genomic studies to researchers, allowing for genome resequencing of hundreds of individuals and in turn improving the identification and taxonomy of wild species, particularly those that are difficult to sample. Furthermore, tissues sampled by minimal or non-invasive methods cannot be identified morphologically and an efficient method for species identification will benefit the sample pre-treatment and selection for large-scale genome resequencing studies.

Current HTS based methods for DNA barcoding are not only cost prohibitive, but are also limited in read length or require extra laboratory workloads. For example, a maximum read length of 300 bp is available on Illumina's MiSeq platform and only delivers a fraction of the standard barcode [17], while multiple rounds of PCRs [18, 19] or an extra K-mer based assembly step (SOAPBarcode [20]) increases laboratory work and leads to accuracy uncertainty [21] (Fig. 1a). Although long reads from the Single Molecular Real Time (SMRT) sequencing platform or nanopore platform can achieve reliable standard barcode sequences, these are at a higher cost than those HTS based methods [21, 22]. Since a standard DNA barcode (e.g. *COI*) with flanking primers and tags can reach ca. 700 bp in length, the HTS platform offers significant advantages provided it can generate reads of ≥400 bp in length, thus forming a minimum overlap of ~ 80 bp (Fig. 1b), which will allow for accurate *COI* barcode assembly by means of simply connecting the 5′ and 3′ reads.

The MGISEQ platform utilizes a technology called DNBSEQ (https://en.mgitech.cn/products/), which amplifies small fragments of genomic DNA into DNA nanoballs by rolling circle amplification, and determines the DNA nanoballs' sequence using a refined combinatorial Probe Anchor Synthesis (cPAS) sequencing technology [23]. It generates sequences in FASTQ format with quality scores based on a Phred-33 standard (equivalent to Illumina's NovaSeq system). Several studies have validated its sequencing quality by comparing its performance with that of Illumina generated sequence data from ancient DNA [24], whole-genome [25] and metagenome

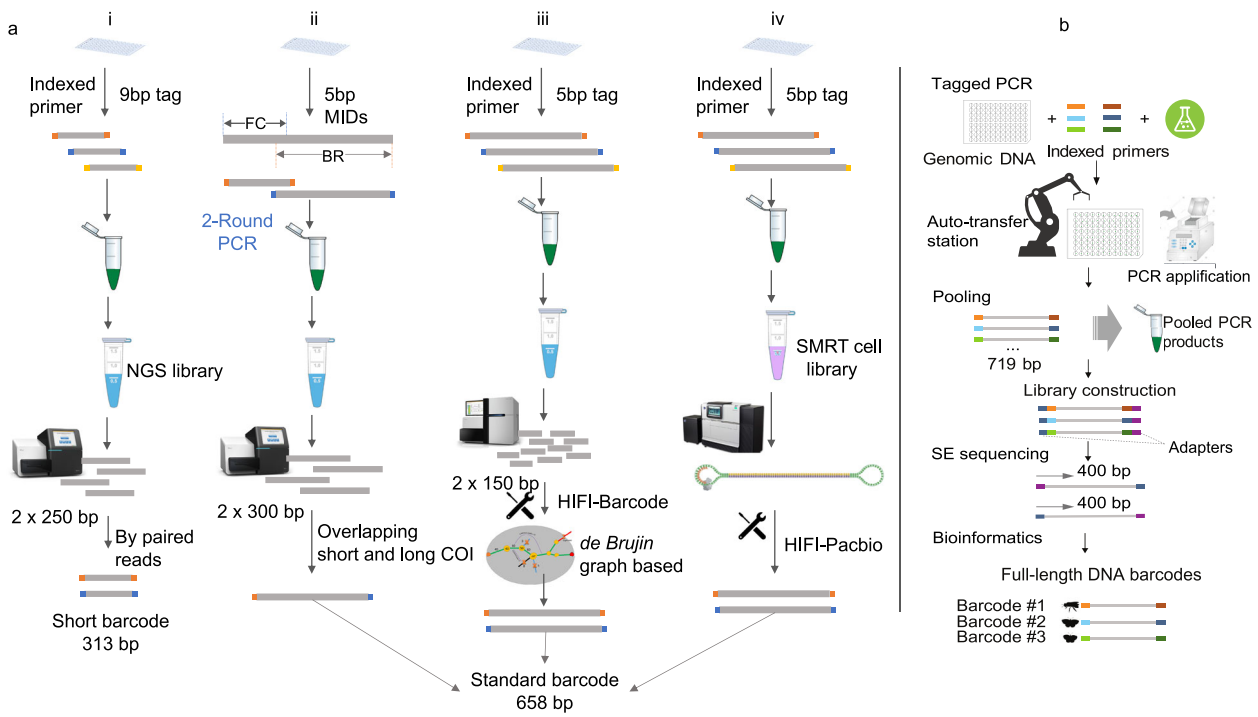

Fig. 1 Comparison of different strategies to access *COI* barcode using HTS platforms. The different experimental designs and adopted sequencing strategies fit for sequencing length capacity (**a**). For four main methods of previous studies, (i) and (ii) refer to (Meier, Wong, Srivathsan, & Foo, 2016), (Shokralla et al., 2015), respectively, while (iii) and (iv) refers to (Liu, Yang, Zhou, & Zhou, 2017). The HIFI-SE pipeline can easily and directly obtain the standard *COI* barcode by overlapping single-end 400 bp (**b**)

sample types [26]. The MGISEQ platform has launched a new sequencing kit capable of single-end 400 bp sequencing - SE400 [27], which offers a simple and reliable way to achieve DNA barcodes efficiently. In this study, we explore the potential of the MGISEQ SE400 sequencing in DNA barcode reference construction and quick species identification, and provide an updated HIFI-SE barcode software package that can generate *COI* barcode assemblies using HTS reads of 400 bp length.

Results

A total of 73 out of 96 (78%, excluding 2 blanks) samples were successfully sequenced and assembled using Sanger sequencing, with the 21 failed samples referred to as "Barcode failed" samples. Comparatively, for the same 96 samples our pipeline produced a total of 12,745,067 HTS SE400 reads that were retained after quality control and around 77.9% (9,870,823) of reads were assigned to their corresponding samples at either the 5′ or 3′ end. The number of sequences of each sample varied markedly, ranging from 303 to 585,609, with Sanger "barcode failed" samples possessing a lower but insignificant number of reads (Additional file 1: Figure S1). Overall, 86 barcode sequences including 63 insect samples and 23 marine invertebrate samples were achieved using the HIFI-SE pipeline, with 14 out of the 21 Sanger "barcode failed" samples being successfully recovered, leading to an overall success rate of 91.5% (Fig. 2). Conversely, one sample that had a Sanger reference did not successfully assemble using our HIFI-SE pipeline. For the remaining samples, an average of 2,457,295 reads per plate were generated and the output profile and successful assignment ratio were on par with that of Plate #1, producing a total of 661 full-length *COI* barcodes (Additional file 2: Table S3).

When comparing our HIFI-SE assembled sequences to the Sanger reference sequences (72 sequences available), HIFI-SE assemblies showed a high-similarity score for the vast majority of the samples (71/72), including 46 samples that had a sequence similarity of 100% and 25 of ~ 99% (Additional file 2: Table S4). Only one sample displayed a high dissimilarity score to its corresponding Sanger reference sequence. A further examination discovered that its sequence was identical to that of another sample on the same plate, so could have been contaminated by that sample. Read alignment showed that the sites on HIFI-SE assemblies at which mismatches occurred were supported by high read coverage, confirming the accurate recovery of HIFI-SE assemblies (Additional file 1: Figure S2). In addition, HIFI-SE identified a total of

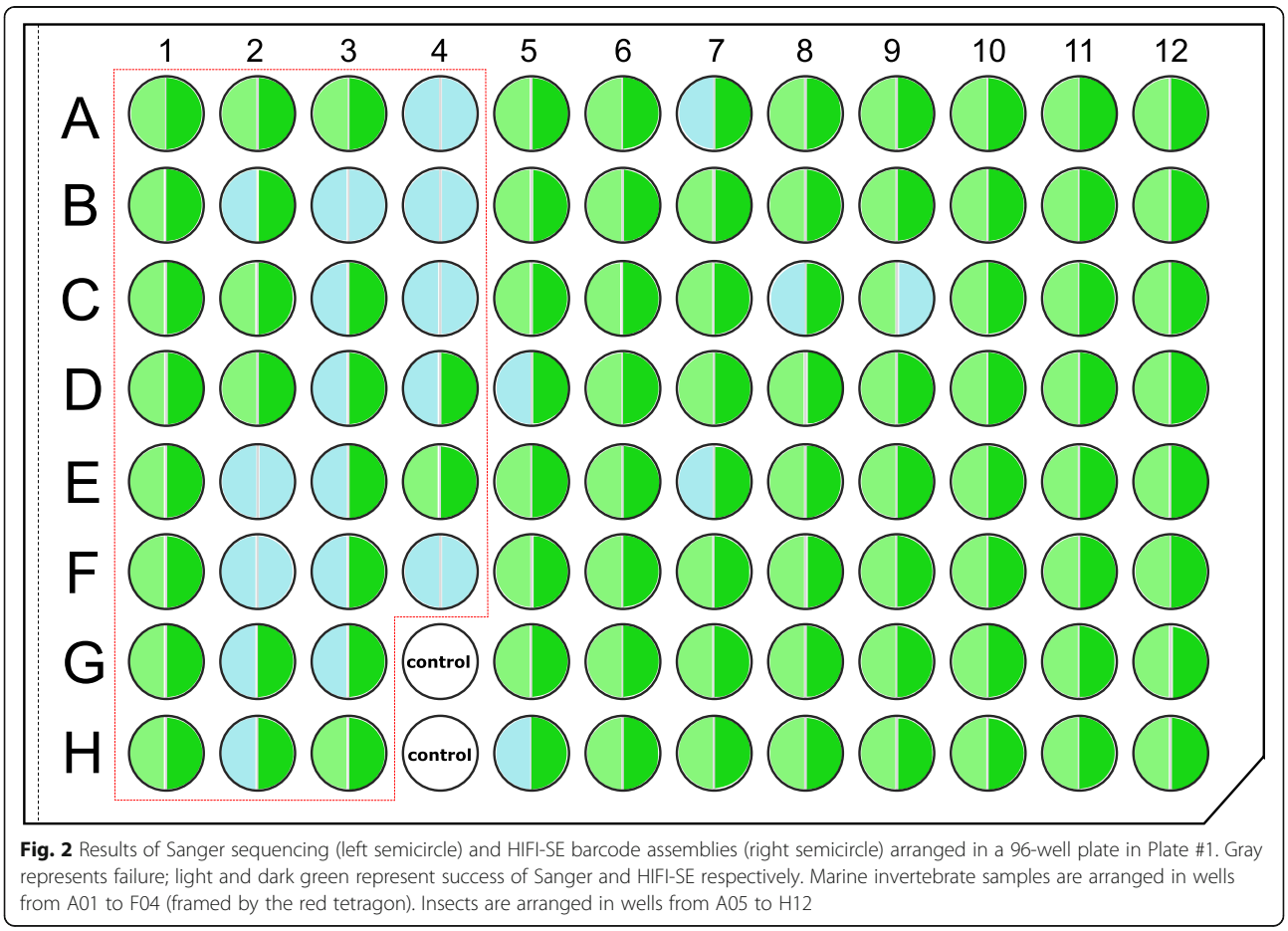

Fig. 2 Results of Sanger sequencing (left semicircle) and HIFI-SE barcode assemblies (right semicircle) arranged in a 96-well plate in Plate #1. Gray represents failure; light and dark green represent success of Sanger and HIFI-SE respectively. Marine invertebrate samples are arranged in wells from A01 to F04 (framed by the red tetragon). Insects are arranged in wells from A05 to H12

40 ambiguous sites in the Sanger references to specific nucleotides and revealed the heteroplasmy states in some samples (Additional file 1: Figure S2).

For the samples without Sanger references, we first conducted a molecular based taxonomic identification by searching their highly similar records on the BOLD system using the HIFI-SE "*taxonomy*" subprogram. The BOLD database search resulted in a total of 418 samples finding their best hits with similarity scores ≥98% [28–30] and the remaining 243 samples with their best hits with similarity scores ranging from 91.4 to 98% [31, 32]. These sequences represented 21 families of Lepidoptera and an unexpected *Homo sapiens* match (99.86% sequence identity on NCBI), which is likely contamination during wet-lab experiments. To further evaluate the accuracy of the HIFI-SE pipeline, we randomly selected 100 samples which had high-quality photos to identify them morphologically, and then check the conformities between the molecular and morphological identification. For the 91 individuals that successfully produced *COI* barcodes, five records conflicted between the morphological and molecular identification, with the remaining samples being congruent between the two identification approaches (Additional file 1: Figure S3). Since the sequence clusters are supported by many reads, it is possible those taxonomic conflicts resulted from incorrect taxonomic annotations in the BOLD system (Fig. 3 & Additional file 5).

We detected *Wolbachia* derived sequences in 13 samples and fungi derived sequences in 57 samples, including four *Wolbachia* species and 42 fungi species with highly similar records (> 98%) on the BOLD database (Additional file 2: Table S5).

Discussion

Despite the importance of biodiversity in ecosystem functioning [33], global biodiversity continues to be lost

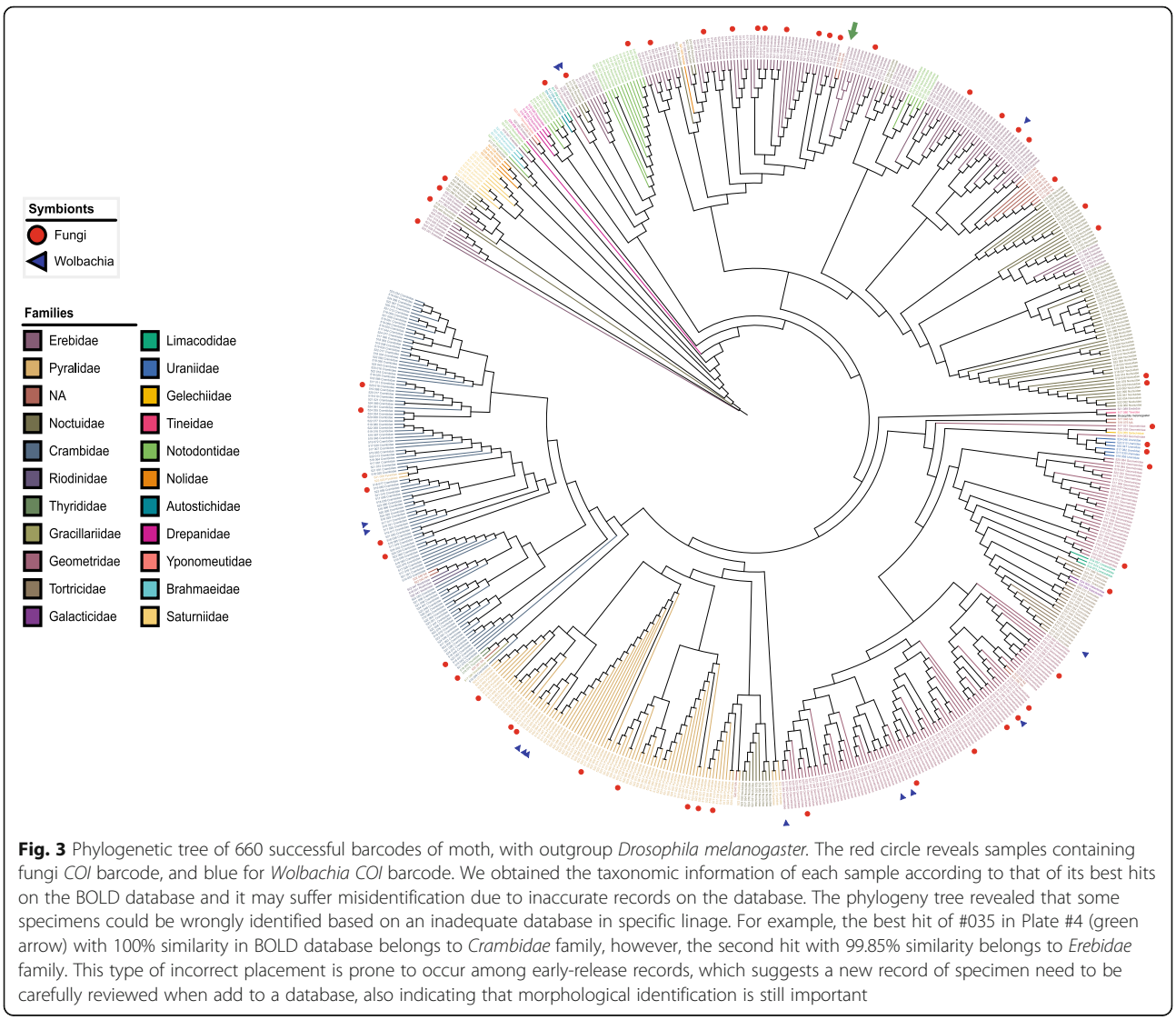

Fig. 3 Phylogenetic tree of 660 successful barcodes of moth, with outgroup *Drosophila melanogaster*. The red circle reveals samples containing fungi *COI* barcode, and blue for *Wolbachia COI* barcode. We obtained the taxonomic information of each sample according to that of its best hits on the BOLD database and it may suffer misidentification due to inaccurate records on the database. The phylogeny tree revealed that some specimens could be wrongly identified based on an inadequate database in specific linage. For example, the best hit of #035 in Plate #4 (green arrow) with 100% similarity in BOLD database belongs to *Crambidae* family, however, the second hit with 99.85% similarity belongs to *Erebidae* family. This type of incorrect placement is prone to occur among early-release records, which suggests a new record of specimen need to be carefully reviewed when add to a database, also indicating that morphological identification is still important

at an unprecedented rate due to climate change and human activities [34]. DNA barcoding has proven effective in accelerating the collection of biodiversity inventories over large geographic and temporal scales, which benefit both researchers and also policy-makers focused on maintaining functioning ecosystems [35]. Burgeoning massive parallel sequencing techniques have driven the cost per nucleotide base down dramatically [36] and facilitated multifaceted approaches to obtain barcode sequences via HTS platforms [20–22]. This has made it possible to generate large amounts of barcode sequences for a tiny fraction of the cost compared to 15 years ago [33, 34, 37].

The HIFI-SE pipeline, that takes advantage of MGIS EQ SE400 reads as long as 400 bp, provides an easy, simple and cost-efficient approach to generate barcode sequences from a large number of samples. The 400 bp reads enable an overlap length of ca. 80 bp for most animal *COI* barcode sequences by sequencing both 5' and 3′ ends. This assembly-by-overlapping step can simplify the barcode assembly process by circumventing the *de Brujin* graph algorithm, which is time-consuming and computationally intensive [38] and can be subject to erroneous pathing when dealing with intricate scenarios.

Currently, high-throughput sequencing platforms (BGI's MGISEQ/T7 or Illumina's HiSEQ/NovaSeq) still have advantages in throughput as well as the cost per base/read over the third-generation platforms (PacBio's SEQUEL II or Oxford Nanopore Technologies' MinION), and the simplified analysis pipeline based on SE400 sequencing is a further advantage. For example, MGISEQ provides a quote of $650 per lane that can produce ca. 275 million reads compared to a quote of $2000 per cell that can produce < 8 million reads with the PacBio's latest SEQUEL II release [39]. However, the third-generation platforms have dramatically increased their sequencing throughput in the last 2 years [39] which, together with its advantage of read length, may surpass the next-generation platforms in barcoding related applications using long fragments (e.g. 16S rRNA gene for bacteria). Similarly, ONT's MinION, a portable and real-time sequencer, can greatly benefit DNA barcoding in terms of speed and flexibility [40]. Thus, while next generation technology is still advantageous for barcoding, third-generation platforms will likely provide useful alternatives in future scenarios.

Two taxonomic groups, marine invertebrates and insects, were sampled in this study to demonstrate the effectiveness of the HIFI-SE approach. The results showed that insects delivered higher barcode recovery ratios (724 out of 815 DNA samples) compared to marine invertebrates (23 out of 30 DNA samples). The relatively lower efficiency of marine invertebrates can be attributed to the biased performance of primer set LCO1490 and HCO2198 [41, 42]. It shows the necessity to improve primer design to cover various phylogenetic lineages in spite of the high sensitivity of HTS methods [43]. The primer's inadequacy for marine invertebrates was also reflected by excessive short co-amplicons (400 ~ 500 bp) detected in 16 out of 21 Sanger "Barcode failed" samples (Additional file 1: Figure S1), which might be derived from nuclear-encoded mitochondrial DNA (NuMT, [44, 45]) and in turn affect the recovery success of their barcode sequences via both the Sanger sequencing and the HIFI-SE pipeline. Additionally, coral is well known for being difficult in terms of DNA extraction and genomic DNA tends to degrade quite rapidly for many species [46], further contributing to the short co-amplicons. However, this also reveals the strength of our approach by sequencing those samples that are difficult to work with. In addition, we also noticed one assembly (E08 in Additional file 2: Table S4) that showed low similarity to its corresponding Sanger reference was actually cross contamination from another cell (C11 or H12 in Fig. 2). Since we mixed PCR reagents and PCR products using an auto transfer station (Hamilton Microlab® STAR) and sample E08 only contained a read number of 1000, we believe this contamination event could result from pipette failure on the auto transfer station during sample transfer, and a subsequent tag hopping from other samples during library construction and sequencing.

We also noticed that a relative low ratio (69.64%) of clean reads can successfully be assigned to their corresponding samples (Additional file 2: Table S3). A further examination for those unassigned reads found that around 50.8% of them were attributed to chimeras, with primer sequences occurring at unexpected positions on the reads (not at the end), and 49.2% failed to match the tagged primer set due to containing > 2 mismatches. This high proportion of chimeric sequences could be formed during PCR and can be derived by many factors [47], such as PCR ramp and cycles [48, 49], DNA template [50], and DNA polymerases errors [51]. The dual-index method utilized in the current study was shown to be an efficient way to eliminate those problematic PCR artifacts [52]. In addition, we also included an option for a "*taxonomy*" module in HIFI-SE that can BLAST the 5′ and 3′ end of the barcode sequences and then compare taxonomies for consistency to further validate the assembly accuracy. Furthermore, NuMTs can be easily identified by HiFi-SE because most of them are less than 300 bp [53] and thus contain both the forward and reverse primer on a single read. It is also worth noting that two blank samples retrieved *COI* barcodes using the default parameter settings – minimum read number requirement of 10 - reaching a read support number of 13.5 and 12.5, respectively. Thus, the

parameter setting for the minimum read number support should be adjusted case by case according to the sequencing depth and the read number of the blank samples.

Although approximately 65% of insect species are estimated to harbor *Wolbachia* [54], we merely detected *Wolbachia* in 13 samples out of 751 moth samples. The low detection ratio could result from the DNA extraction strategy and PCR primer biases, so extra primer sets designed for *Wolbachia* may increase the chances to detect symbiotic bacteria. Further, the fungus detected here were all derived from a single phylum Ascomycota, which contains many well-known fungi that infect and kill insects [55, 56], e.g. *Metarhizium anisopliae*, and fungus in genus *Penicillium*. This taxonomic connection is of interest and deserves further investigation to identify the species interactions which is a focus of major research initiatives such as the BIOSCAN project [37, 57] (https://ibol.org/programs/bioscan/).

Conclusion

In summary, the HIFI-SE pipeline requires straightforward processing in both sequencing preparation and data analysis, and holds potential to further reduce per unit cost of DNA barcoding while increasing the efficiency and accuracy of the obtained barcodes. Further cost reduction can be achieved by increasing tag length to allow more index combinations, and pooling amplicons using different primer sets. In addition, although we used the *COI* barcode for demonstration, our pipeline is expected to fit other marker genes with a length of 600-750 bp (e.g. V1-V4, V3-V6, and V5-V9 of 16S rRNA gene). Therefore, this new approach can produce standard full-length barcodes cost efficiently, allowing initiatives targeted at DNA barcoding of different biomes to be more achievable, thereby improving our understanding of the biodiversity of global ecosystems or improving DNA based biosecurity programs. Furthermore, the detection of symbiont information using the current protocol provides an efficient way to study the network and adaptive evolution between the hosts and their symbionts or parasites [58–60].

Methods

Sample collection and DNA extraction

A total of 845 samples, including marine invertebrates (30 samples) and insects (815 samples) were used to test our COI barcoding pipeline (Additional file 2: Table S1). Marine invertebrates were collected using a hammer and chisel (for sceractinian coral) or sterile razor blades (octocorals and sponges) in May 2017, from Orpheus Island in the central in-shore region of the Great Barrier Reef, under the Marine Parks permit G15/37574.1. Coral tissue was removed from the skeleton using pressurized air from a blow gun into a ziplock bag containing 10 mL of calcium magnesium free artificial seawater (CMFA SW; NaCl 26.2 g, KCl 1 g, $NaHCO_3$, Milli-Q H_2O 1 L). Coral tissue blastate was aliquoted into 2 mL microfuge tubes and pelleted in a fixed angle centrifuge at 10,000 x g for 10 min. Pellets were snap frozen and stored at − 80 °C until DNA extraction. All other marine invertebrates were dissected to fit into a 2 mL cryovial, snap frozen and stored at − 80 °C until DNA extraction. Insect samples were collected in August 2017 from the Laohegou Natural Reserve in Sichuan Province and from the Lushan Town, Zhoushan City, Zhejiang Province in China via light trapping. Approximately 0.05 g of coral tissue pellet or marine invertebrate tissue was then used for DNA extraction using the PowerBiofilm DNA Isolation Kit (QIAGEN Pty Ltd., Australia) following the manufacturers protocol. The DNA of insects were extracted using the Glass Fiber Plate method [61], or using the tissue/cell genomic DNA rapid extraction kit (Tiangen Biochemical Technology Co., Ltd., Beijing).

Tag design, PCR amplification, and sanger sequencing

A total of 96 paired tags were added to both ends of the common *COI* barcode primer set (LCO1490 and HCO2198 [62]) (Additional file 2: Table S2). The tag sequence was 5 bp in length and had ≥2 bp difference from each other. Each PCR reaction (25 μL) contained 1 μL DNA template, 16.2 μL molecular biology grade water, 2.5 μL 10× buffer (Mg^{2+} plus), 2.5 μL dNTP mix (10 mM), 1 μL each forward and reverse primers (10 mM), and 0.3 μL TaKaRa Ex Taq polymerase (5 U/μL) (Takara, Dalian, China). The amplification program included a thermocycling profile of 94 °C for 60s, 5 cycles of 94 °C for 30 s, 45 °C for 40 s, and an extension at 72 °C for 60 s, followed by 35 cycles of 94 °C for 30 s, 51 °C for 40 s, and 72 °C for 60 s, with a final extension at 72 °C for 10 min, and a final on-hold at 12 °C. Amplicons generated from the plate (plate #1) containing both the marine invertebrate and insect species were individually visualized on a 1.2% 96 Agarose E-gel (Biowest Agarose) and Sanger sequenced using an ABI 3730XL sequencer (BGI-Shenzhen) and then assembled using Geneious [63].

Library construction and sequencing

One microliter of each amplicon was mixed and sent to BGI-Shenzhen for library preparation and sequencing (MGISEQ SE400 module) following the general library construction protocol (Additional file 3), with a minor modification to exclude DNA fragmentation and size selection.

HIFI-SE: Bioinformatic analysis for SE400 data

To increase accessibility of our newly developed pipeline using the MGISEQ-2000 platform with 400 bp single-end sequencing, we developed a software package, HIFI-SE, which is written in Python and is deposited on PyPI (https://pypi.org/project/HIFI-SE/), consisting of four main function modules of *'filter'*, *'assign'*, *'assembly'* and *'taxonomy'* (Fig. 4). Full function instruction and a tutorial are detailed in the software manual (Additional file 4) and briefly outlined below.

Data filtering

Removes low quality reads including; 1) reads containing any ambiguous bases (i.e. "N") and 2) reads with an expected error number $E^* > 10$ with E^* calculated using a formula of $E^* = \sum_{i=1}^{n} 10^{-Qi/10}$, where n represents sequence length and Qi represents base quality (Phred-33 standard) of the i^{th} base on reads.

Read assignment

Reads were demultiplexed by index and classified to the 5′ and 3′ ends according to the primer sequences, allowing one base mismatch in the index region and one base mismatch in the primer region. In addition, since tagged primer sequences are expected to be located at the end of each read, primer sequences found in improper positions (e.g. in the middle) of the reads were regarded as chimeras and removed automatically during the assignment. Finally, all reads were classified into 192 (96*2) groups consisting of both the 5′ and 3′ end for each of the 96 tags.

Full-length COI barcode assembly

Sequences within each group were first clustered at a 98% similarity using VSEARCH (v2.8.0) [64] and a consensus sequence was built from the most abundant cluster. Additionally, a consensus sequence of the second most abundant cluster was also retained if the number of sequences in that cluster was greater than 1/10 of the top cluster, to identify potential symbionts or parasites. Finally, a minimum sequence number of five for each

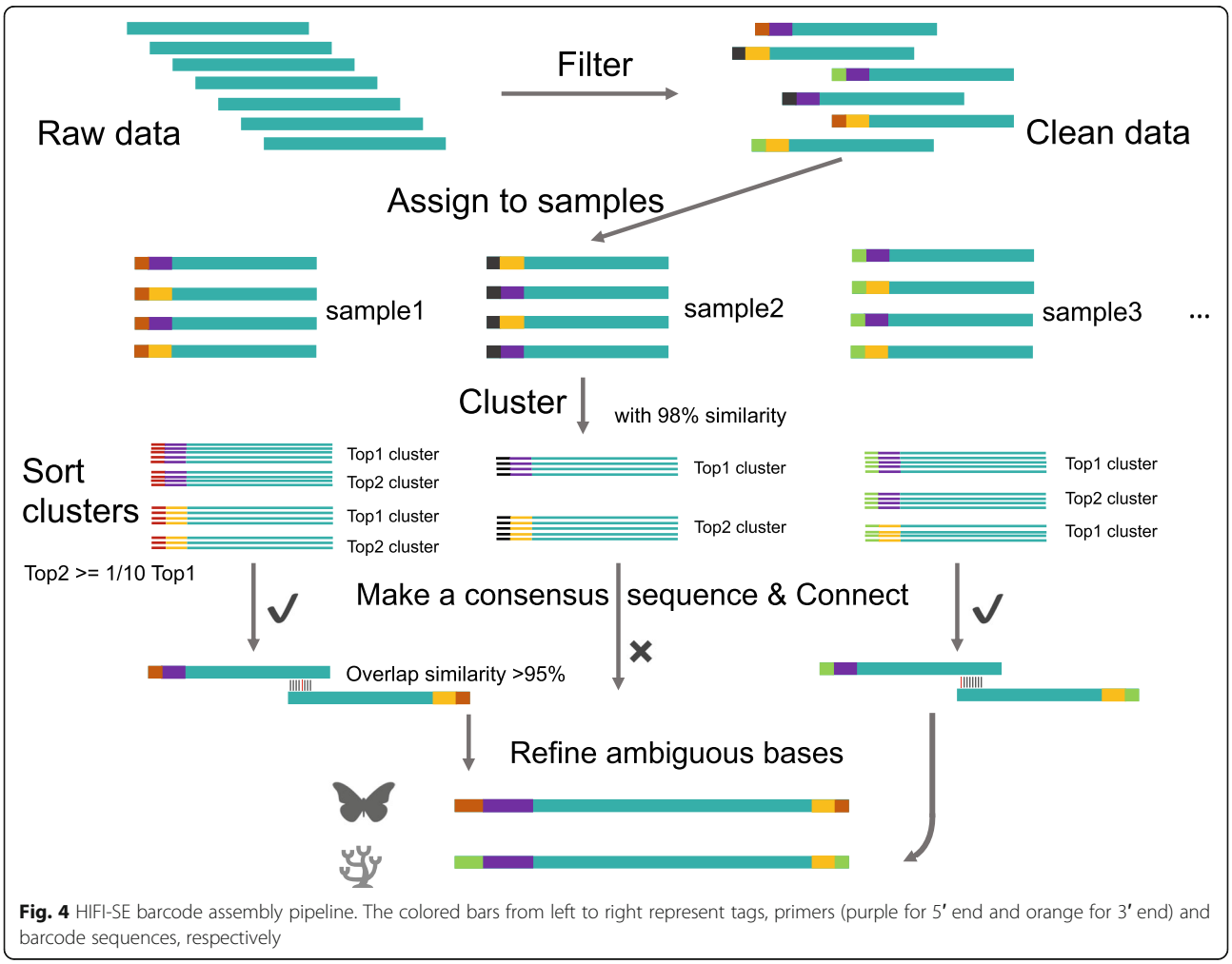

Fig. 4 HIFI-SE barcode assembly pipeline. The colored bars from left to right represent tags, primers (purple for 5′ end and orange for 3′ end) and barcode sequences, respectively

cluster is needed to guarantee the accuracy of the consensus sequence.

Full-length *COI* barcodes were assembled by connecting the consensus sequences of the 5′ and 3′ ends with an overlap ≥80 bp and a similarity ≥95% (mismatches may exist in the overlapping regions due to reduced read quality when towards the read ends). Mismatches in the overlapped region were determined based on the base frequency calculated from sequences in both ends. The assemblies with correct amino acid translation (without stop codons) and a length of > 650 bp were output as the final results. Users also have the flexibility to run another assembly with an additional parameter in the event samples fail with the default parameter settings, for example, by checking for amino acid translation before clustering (Additional file 4).

Taxonomy identification in BOLD

The HIFI-SE pipeline provides an optional step (*taxonomy*) to verify the taxonomic information of the assembled sequences. It can automatically submit assemblies to the BOLD system and retrieve the taxonomic information from the returned searches. Currently, it supports searching of the animal, fungi and plant databases and outputs a user-defined number of BOLD items for each sequence.

Performance evaluation based on the test samples

COI barcode retrieval and symbiont detection

We obtained *COI* barcode assemblies for each sample using the HIFI-SE package with default parameter settings. To further detect nontarget *COI* barcodes (e.g. *Wolbachia* and fungi), all the non-targeted clusters with sequence numbers ≥10 were assembled with default settings. We also identified potential symbionts via BLAST searching [65] (version 2.7.1+, E-value ≤1e-5) a manually curated symbiont dataset (COI genes downloaded from NCBI Genbank, https://github.com/comery/HIFI-barcode-SE400/) before submitting all the barcode assemblies to the BOLD system for taxonomic identification.

Accuracy estimation

For the samples that were Sanger sequenced, we assembled and achieved the barcode sequences using Geneious [63]. To evaluate the accuracy of HIFI-SE pipeline, the HIFI-SE assemblies were aligned to their Sanger references using MUSCLE (v3.8.31) [66] and then checked for similarities between each. We subsequently aligned the demultiplexed reads to their corresponding HIFI-SE assemblies using BWA (Version: 0.7.17-r1188) [67] to examine read support for sites at which the HIFI-SE assemblies and Sanger sequences were different.

Species identification and phylogenetic analysis

Species identification was implemented by HIFI-SE "*taxonomy*" function with a setting of "-n 5 (output five best hits)". We inferred the phylogenetic relationship for all lepidopteran *COI* barcode sequences using IQ-TREE (version 1.6.5) [68] with *Drosophila melanogaster* used as an outgroup after alignment using MAFFT (v7.245) [69] with the parameters of *"--localpair --maxiterate 16 --phylipout --reorder"*.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12864-020-07255-w.

Additional file 1: Figure S1. Read counts of the Sanger barcode failed samples. Stars indicate samples of which short amplicon(s) was detected in the HIFI assemblies. Short amplicons are those clusters of abundance > 10 and of length < 600 bp. The bar plot demonstrates the number of assigned reads for the barcode failed samples. The red dashed line shows the average value of all the successful samples and no significant difference was detected between the two groups (*P* value of 0.232, Student's t-Test). **Figure S2.** Discrepancies between Sanger sequences and HIFI-SE barcodes. Entropy weight was calculated based on the strength of read depth by aligning the SE400 reads onto the assembled HIFI-SE barcodes, showing differences between ambiguous Sanger base-calling and specific nucleotide identified in HIFI-SE barcodes (A) and potential heteroplasmy (B). In addition, several N bases were present of insertion in Sanger sequence (C), also two N bases in HIFI sequences (D). **Figure S3.** Comparison of molecular and morphological identification.

Additional file 2: Table S1. Sequence of the tagged primers. **Table S2.** Sample Information. **Table S3.** Statistical results of data output and *COI* barcode recovery. **Table S4.** Accuracy results of HIFI-SE barcodes compared with Sanger. **Table S5.** *Wolbachia* and fungi sequences detected from moth samples.

Additional file 3. Library construction protocol of MGISEQ-2000 SE400 module.

Additional file 4. The manual of HIFI-SE package.

Additional file 5. A note for taxonomy identification issue of sample #035 in plate #4.

Abbreviations

SE400: Single-end 400 bp long read sequencing; *COI*: Cytochrome *c* oxidase subunit I; BOLD: The Barcode of Life Data systems; HTS: High-throughput sequencing; SMRT: Single Molecular Real Time; HIFI-SE: High-throughput and high-fidelity DNA barcoding analysis pipeline for single-end 400 bp long reads

Acknowledgements

We thank Dr. Ding Yang from China Agricultural University for contributing samples. We would like to thank Guojie Zhang and Qiye Li for sample and SE400 sequencing coordination. Field work at Orpheus Island was support by EarthWatch Australia. Samples from Orpheus Island were collected under the Marine Parks permit G15/37574.1.

Authors' contributions

C.Y. and SL.L. conceived the idea and designed the methodology; C.Y. and G.M. developed the program; D.G.B and P.A.O. collected the marine invertebrate samples and extracted DNA. Y.Z. and C.Y. collected moth samples. C.Y. and S.T. collected, analyzed the data and drafted the manuscript; J.X., SH.L., A.C. and X.C. conducted the library construction and sequencing. S.L. and A.Z. revised the manuscript. All authors have read and approved the manuscript.

Funding

This research was supported by Chinese Postdoctoral Science Foundation (2019 M660051), Shenzhen Municipal Government of China (NO. JCYJ20170817150755701),

Shenzhen Peacock Plan (No. KQTD20150330171505310), and partly supported by China National Funds for Distinguished Young Scientists [grant number 31425023].

Availability of data and materials
The datasets generated and/or analyzed during the current study are available in the CNGB Nucleotide Sequence Archive (CNSA: https://db.cngb.org/cnsa; accession number CNP0000195, and the EMBL repository (PRJEB29212, ERP111495). The HIFI-SE program and symbiont dataset used in this study were deposited on Github (https://github.com/comery/HIFI-barcode-SE400).

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
Authors have no conflict of interest to declare.

Author details
[1]BGI-Shenzhen, Shenzhen 518083, China. [2]College of Life Sciences, Capital Normal University, Beijing 100048, China. [3]College of Science and Engineering, James Cook University, Townsville, QLD, Australia. [4]Australian Institute of Marine Science, Townsville, QLD, Australia. [5]AIMS@JCU, Townsville, QLD, Australia. [6]Beijing Advanced Innovation Center for Food Nutrition and Human Health, College of Plant Protection, China Agricultural University, Beijing 100193, China.

Received: 23 December 2019 Accepted: 18 November 2020
Published online: 04 December 2020

References
1. Hebert PD, Cywinska A, Ball SL. Biological identifications through DNA barcodes. Proc R Soc Lond B Biol Sci. 2003;270(1512):313–21.
2. Ratnasingham S, Hebert PD. BOLD: the barcode of life data systemhttp://www.barcodinglife.org. Mol Ecol Notes. 2007;7(3):355–64.
3. Valentini A, Pompanon F, Taberlet P. DNA barcoding for ecologists. Trends Ecol Evol. 2009;24(2):110–7.
4. Symondson WO. Molecular identification of prey in predator diets. Mol Ecol. 2002;11(4):627–41.
5. Krehenwinkel H, Kennedy S, Pekár S, Gillespie RG. A cost-efficient and simple protocol to enrich prey DNA from extractions of predatory arthropods for large-scale gut content analysis by Illumina sequencing. Methods Ecol Evol. 2017;8(1):126–34.
6. Bączkiewicz A, Szczecińska M, Sawicki J, Stebel A, Buczkowska K. DNA barcoding, ecology and geography of the cryptic species of Aneura pinguis and their relationships with Aneura maxima and Aneura mirabilis (Metzgeriales, Marchantiophyta). PLoS One. 2017;12(12):e0188837.
7. Tang M, Hardman CJ, Ji Y, Meng G, Liu S, Tan M, et al. High-throughput monitoring of wild bee diversity and abundance via mitogenomics. Methods Ecol Evol. 2015;6(9):1034–43.
8. Thomsen PF, Willerslev E. Environmental DNA – an emerging tool in conservation for monitoring past and present biodiversity. Biol Conserv. 2015;183:4–18.
9. Bohmann K, Evans A, Gilbert MTP, Carvalho GR, Creer S, Knapp M, et al. Environmental DNA for wildlife biology and biodiversity monitoring. Trends Ecol Evol. 2014;29(6):358–67.
10. Krishnamurthy PK, Francis RA. A critical review on the utility of DNA barcoding in biodiversity conservation. Biodivers Conserv. 2012;21(8):1901–19.
11. Collins RA, Armstrong KF, Meier R, Yi Y, Brown SD, Cruickshank RH, et al. Barcoding and border biosecurity: identifying cyprinid fishes in the aquarium trade. PLoS One. 2012;7(1):e28381.
12. Coghlan ML, Haile J, Houston J, Murray DC, White NE, Moolhuijzen P, et al. Deep sequencing of plant and animal DNA contained within traditional Chinese medicines reveals legality issues and health safety concerns. PLoS Genet. 2012;8(4):e1002657.
13. Yoccoz NG. The future of environmental DNA in ecology. Mol Ecol. 2012;21(8):2031–8.
14. Clarke LJ, Soubrier J, Weyrich LS, Cooper A. Environmental metabarcodes for insects: in silico PCR reveals potential for taxonomic bias. Mol Ecol Resour. 2014;14(6):1160–70.
15. Curry CJ, Gibson JF, Shokralla S, Hajibabaei M, Baird DJ. Identifying north American freshwater invertebrates using DNA barcodes: are existing COI sequence libraries fit for purpose? Freshwater Sci. 2018;37(1):178–89.
16. Porter TM, Hajibabaei M. Over 2.5 million COI sequences in GenBank and growing. PLoS One. 2018;13(9):e0200177.
17. Meier R, Wong W, Srivathsan A, Foo M. $1 DNA barcodes for reconstructing complex phenomes and finding rare species in specimen-rich samples. Cladistics. 2016;32(1):100–10.
18. Shokralla S, Porter TM, Gibson JF, Dobosz R, Janzen DH, Hallwachs W, et al. Massively parallel multiplex DNA sequencing for specimen identification using an Illumina MiSeq platform. Sci Rep. 2015;5:9687.
19. Cruaud P, Rasplus JY, Rodriguez LJ, Cruaud A. High-throughput sequencing of multiple amplicons for barcoding and integrative taxonomy. Sci Rep. 2017;7:41948.
20. Liu S, Li Y, Lu J, Su X, Tang M, Zhang R, et al. SOAPBarcode: revealing arthropod biodiversity through assembly of Illumina shotgun sequences of PCR amplicons. Methods Ecol Evol. 2013;4(12):1142–50.
21. Liu S, Yang C, Zhou C, Zhou X. Filling reference gaps via assembling DNA barcodes using high-throughput sequencing—moving toward barcoding the world. GigaScience. 2017;6(12):1–8.
22. Hebert PD, Braukmann TW, Prosser SW, Ratnasingham S, Ivanova NV, Janzen DH, et al. A Sequel to sanger: amplicon sequencing that scales. BMC Genomics. 2018;19(1):219.
23. Drmanac R, Sparks AB, Callow MJ, Halpern AL, Burns NL, Kermani BG, et al. Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. Science. 2010;327(5961):78–81.
24. Mak SST, Gopalakrishnan S, Carøe C, Geng C, Liu S, Sinding M-HS, et al. Comparative performance of the BGISEQ-500 vs Illumina HiSeq2500 sequencing platforms for palaeogenomic sequencing. Gigascience. 2017;6(8):gix049.
25. Korostin D, Kulemin N, Naumov V, Belova V, Kwon D, Gorbachev A. Comparative analysis of novel MGISEQ-2000 sequencing platform vs Illumina HiSeq 2500 for whole-genome sequencing. PLoS One. 2020;15(3):e0230301.
26. Fang C, Zhong H, Lin Y, Chen B, Han M, Ren H, et al. Assessment of the cPAS-based BGISEQ-500 platform for metagenomic sequencing. Gigascience. 2018;7(3):gix133.
27. Longer read length, wider application - MGISEQ-2000RS high-throughput sequencing reagent kit now available (SE400)-MGI-MGI Tech Co., Ltd. [https://en.mgitech.cn/article/detail/SE400.html].
28. Hebert PDN, Ratnasingham S, De Waard JR. Barcoding animal life: cytochrome c oxidase subunit 1 divergences among closely related species. Proc R Soc Lond Ser B. 2003;270(suppl_1):S96–9.
29. Zhou X, Adamowicz SJ, Jacobus LM, DeWalt RE, Hebert PDN. Towards a comprehensive barcode library for arctic life-Ephemeroptera, Plecoptera, and Trichoptera of Churchill, Manitoba, Canada. Front Zool. 2009;6(1):30.
30. Ruiter DE, Boyle EE, Zhou X. DNA barcoding facilitates associations and diagnoses for Trichoptera larvae of the Churchill (Manitoba, Canada) area. BMC Ecol. 2013;13(1):5.
31. Park D-S, Foottit R, Maw E, Hebert PDN. Barcoding bugs: DNA-based identification of the true bugs (Insecta: Hemiptera: Heteroptera). PLoS One. 2011;6(4):e18749.
32. Kim J, Jung S. COI barcoding of plant bugs (Insecta: Hemiptera: Miridae). PeerJ. 2018;6:e6070.
33. Tilman D, Isbell F, Cowles JM. Biodiversity and ecosystem functioning. Annu Rev Ecol Evol Syst. 2014;45:471–93.
34. Kerr JT, Currie DJ. Effects of human activity on global extinction risk. Conserv Biol. 1995;9(6):1528–38.
35. Weigand H, Beermann AJ, Ciampor F, Costa FO, Csabai Z, Duarte S, et al. DNA barcode reference libraries for the monitoring of aquatic biota in Europe: gap-analysis and recommendations for future work. Sci Total Environ. 2019;678:499–524.
36. Von Bubnoff A. Next-generation sequencing: the race is on. Cell. 2008;132(5):721–3.
37. Hobern D. BIOSCAN: DNA barcoding to accelerate taxonomy and biogeography for conservation and sustainability. Genome. 2020;999:1–4.

38. Li Z, Chen Y, Mu D, Yuan J, Shi Y, Zhang H, et al. Comparison of the two major classes of assembly algorithms: overlap–layout–consensus and de-bruijn-graph. Brief Funct Genom. 2012;11(1):25–37.
39. Logsdon GA, Vollger MR, Eichler EE. Long-read human genome sequencing and its applications. Nat Rev Genet. 2020;21(10):597.
40. Menegon M, Cantaloni C, Rodriguez-Prieto A, Centomo C, Abdelfattah A, Rossato M, et al. On site DNA barcoding by nanopore sequencing. PLoS One. 2017;12(10):e0184741.
41. Geller J, Meyer C, Parker M, Hawk H. Redesign of PCR primers for mitochondrial cytochrome c oxidase subunit I for marine invertebrates and application in all-taxa biotic surveys. Mol Ecol Resour. 2013;13(5):851–61.
42. Deagle BE, Jarman SN, Coissac E, Pompanon F, Taberlet P. DNA metabarcoding and the cytochrome c oxidase subunit I marker: not a perfect match. Biol Lett. 2014;10(9):20140562.
43. Elbrecht V, Braukmann TWA, Ivanova NV, Prosser SWJ, Hajibabaei M, Wright M, et al. Validation of COI metabarcoding primers for terrestrial arthropods. PeerJ. 2019;7:e7745.
44. Sorenson MD, Quinn TW. Numts: a challenge for avian systematics and population biology. Auk. 1998;115(1):214–21.
45. Holekamp KE, Sakai ST, Lundrigan BL. Social intelligence in the spotted hyena (Crocuta crocuta). Philos Trans R Soc B. 2007;362(1480):523–38.
46. Neigel J, Domingo A, Stake J. DNA barcoding as a tool for coral reef conservation. Coral Reefs. 2007;26(3):487.
47. Haas BJ, Gevers D, Earl AM, Feldgarden M, Ward DV, Giannoukos G, et al. Chimeric 16S rRNA sequence formation and detection in sanger and 454-pyrosequenced PCR amplicons. Genome Res. 2011;21(3):494–504.
48. Stevens JL, Jackson RL, Olson JB. Slowing PCR ramp speed reduces chimera formation from environmental samples. J Microbiol Methods. 2013;93(3):203–5.
49. Vierna J, Dona J, Vizcaino A, Serrano D, Jovani R. PCR cycles above routine numbers do not compromise high-throughput DNA barcoding results. Genome. 2017;60(10):868–73.
50. Kalle E, Kubista M, Rensing C. Multi-template polymerase chain reaction. Biomol Detect Quantification. 2014;2:11–29.
51. Qiu X, Wu L, Huang H, McDonel PE, Palumbo AV, Tiedje JM, et al. Evaluation of PCR-generated chimeras, mutations, and heteroduplexes with 16S rRNA gene-based cloning. Appl Environ Microbiol. 2001;67(2):880–7.
52. Kozich JJ, Westcott SL, Baxter NT, Highlander SK, Schloss PD. Development of a dual-index sequencing strategy and curation pipeline for analyzing amplicon sequence data on the MiSeq Illumina sequencing platform. Appl Environ Microbiol. 2013;79(17):5112–20.
53. Richly E, Leister D. NUMTs in sequenced eukaryotic genomes. Mol Biol Evol. 2004;21(6):1081–4.
54. Hilgenboecker K, Hammerstein P, Schlattmann P, Telschow A, Werren JH. How many species are infected with Wolbachia?–a statistical analysis of current data. FEMS Microbiol Lett. 2008;281(2):215–20.
55. Shang Y, Feng P, Wang C. Fungi that infect insects: altering host behavior and beyond. PLoS Pathog. 2015;11(8):e1005037.
56. Mora MAE, Castilho AMC, Fraga ME. Classification and infection mechanism of entomopathogenic fungi. Arquivos do Instituto Biológico. 2017;84:1.
57. Hobern D, Hebert P. BIOSCAN-revealing eukaryote diversity, dynamics, and interactions. Biodivers Inform Sci Standards. 2019;3:e37333.
58. Dobson SL, Fox CW, Jiggins FM. The effect of Wolbachia-induced cytoplasmic incompatibility on host population size in natural and manipulated systems. Proc R Soc Lond Ser B Biol Sci. 2002;269(1490):437–45.
59. Zabalou S, Riegler M, Theodorakopoulou M, Stauffer C, Savakis C, Bourtzis K. Wolbachia-induced cytoplasmic incompatibility as a means for insect pest population control. Proc Natl Acad Sci. 2004;101(42):15042–5.
60. Xi Z, Khoo CCH, Dobson SL. Wolbachia establishment and invasion in an Aedes aegypti laboratory population. Science. 2005;310(5746):326–8.
61. Ivanova NV, Dewaard JR, Hebert PD. An inexpensive, automation-friendly protocol for recovering high-quality DNA. Mol Ecol Notes. 2006;6(4):998–1002.
62. MB OF, Hoeh W, Lutz R, Vrijenhoek R. DNA primers for amplification of mitochondrial cytochrome c oxidase subunit I from diverse metazoan invertebrates. Mol Mar Biol Biotechnol. 1994;3(5):294–9.
63. Kearse M, Moir R, Wilson A, Stones-Havas S, Cheung M, Sturrock S, et al. Geneious basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics. 2012;28(12):1647–9.
64. Rognes T, Flouri T, Nichols B, Quince C, Mahé F. VSEARCH: a versatile open source tool for metagenomics. PeerJ. 2016;4:e2584.
65. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: architecture and applications. BMC Bioinform. 2009;10(1):421.
66. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 2004;32(5):1792–7.
67. Li H, Durbin R. Fast and accurate short read alignment with burrows–wheeler transform. Bioinformatics. 2009;25(14):1754–60.
68. Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol Biol Evol. 2014;32(1):268–74.
69. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol. 2013;30(4):772–80.

Publisher's Note