

GigaScience

Filling reference gaps via assembling DNA barcodes using high-throughput sequencing - moving toward barcoding the world --Manuscript Draft--

Manuscript Number:	GIGA-D-17-00172R1	
Full Title:	Filling reference gaps via assembling DNA barcodes using high-throughput sequencing - moving toward barcoding the world	
Article Type:	Research	
Funding Information:	the Chinese Universities Scientific Fund (2017QC114)	Not applicable
Abstract:	<p>Over the past decade, biodiversity scientists have dedicated tremendous efforts in constructing DNA reference barcodes for rapid species registration and identification. Although analytical cost for standard DNA barcoding has been significantly reduced since early 2,000, further dramatic reduction on barcoding costs is unlikely because the Sanger sequencing is approaching its limits in throughput and chemistry cost. Constraints in barcoding cost not only led to unbalanced barcoding efforts around the globe, but also refrained High-Throughput-Sequencing (HTS) based taxonomic identification from applying binomial species names, which provide crucial linkages to biological knowledge. We developed an Illumina-based pipeline, HIFI-Barcode, to produce full-length COI barcodes from pooled PCR amplicons generated by individual specimens. The new pipeline generated accurate barcode sequences that were comparable to Sanger standards, even for different haplotypes of the same species that were only a few nucleotides different from each other. Additionally, the new pipeline was much more sensitive in recovering amplicons at low quantity. The HIFI-Barcode pipeline successfully recovered barcodes from over 78% of the PCR reactions that didn't show clear bands on the electrophoresis gel. Moreover, sequencing results based on the single molecular sequencing platform, Pacbio, confirmed the accuracy the HIFI-Barcode results. Altogether, the new pipeline can provide an improved solution to produce full-length reference barcodes at about 1/10 of the current cost, enabling construction of comprehensive barcode libraries for local fauna, leading to a feasible direction for DNA barcoding global biomes.</p>	
Corresponding Author:	Xin Zhou CHINA	
Corresponding Author Secondary Information:		
Corresponding Author's Institution:		
Corresponding Author's Secondary Institution:		
First Author:	Shanlin Liu	
First Author Secondary Information:		
Order of Authors:	Shanlin Liu	
	Chentao Yang	
	Chengran Zhou	
	Xin Zhou	
Order of Authors Secondary Information:		
Response to Reviewers:	<p>Dear Scott,</p> <p>We have carefully revised the manuscript in accordance with reviewers' comments. Enclosed please find our corresponding responses in detail. We have addressed all issues raised by the reviewers. Thanks to the constructive suggestions from both reviewers, we think our manuscript has been significantly improved. We sincerely hope</p>	

this manuscript is now suitable for publication with GigaScience. Thank you for all your help and looking forward to hearing from you soon.

Sincerely yours,
Xin Zhou

Please find the following responses to editorial comments:

Reviewer comments:

Reviewer #1: The manuscript describes a novel method to recover full COI barcodes for individual specimens using high throughput sequencing technology for construction of reference databases linking species names and the barcode sequences, which are crucial for assessments of biodiversity based on molecular methods like metabarcoding.

The method overcomes a problem of Illumina platform (i.e. length of reads [typically 150+150 bp] is short for standard COI barcode [>600 bp]). Compared with previously published method for the same purpose (ref. 21 Shokralla et al. 2015), the method was improved in cost effectiveness, thanks to the simplified laboratory protocol and superb bioinformatics procedures, although one of the basic ideas (gap-filling) were already published (ref. 23). Further, the authors validated their method by using Pacbio, another (expensive) sequencing platform.

The paper will contribute to studies on biodiversity and meets the scope of the journal. I recommend it be accepted for publication after minor revision.

- Please clarify in the figure legend that sequence logos shown in Fig 4 were based on alignments of Illumina raw reads onto assembled HIFI-barcodes.

Response: Thanks for your comments. We have modified the legend of Fig. 4 to supplement the alignment info – “Entropy weight was calculated based on the strength of read depth by aligning Illumina raw reads onto assembled HIFI-barcodes”.

- What does "heterozygote" mean for mitochondrial genes? Do you mean heteroplasmy?

Response: Thank you for your kindly remind. We have changed it to heteroplasmy.

- I would suggest the authors talk potential co-amplification of nuclear mitochondrial pseudogenes (so-called numts) by PCR.

Response: We appreciate your suggestion for the pseudogenes which has been pointed out in line 5, page 11 – “as well as potential PCR errors and pseudo-genes (numts)”. We have also added another sentence in line 8, page 11, - “therefore ruling out the likelihood of assembly errors in the HIFI-Barcode pipeline, suggesting they are co-amplified numts present in PCR products”

- In the bioinformatics pipeline, to my understanding, reads that were used for gap-filling had no information about from which sample they were derived; Gaps were filled by using de Bruijn graphs which were constructed based on pooled reads (among all samples). If many conspecific specimens are sequenced simultaneously and they harbor SNPs in these gap regions, what happens? I am afraid that the most abundantly sequenced nucleotide type may "overwrite" sequences for the other conspecifics. Please talk on this concern.

Response: Thanks for pointing this out. Yes, according to the algorithm of de Bruijn graphs, single nucleotide variants (SNVs), if locate in the gap filling region, which is, in our COI case, around 200 bp long, can potentially be overwritten by the most abundant genotypes. But correct assembly can be achieved by using large Kmers (our current algorithm applies $Kmer \geq 100$) in the case of haplotypes. In fact, samples A11 and A12 in plate2 showed a proper example: where A11 and A12 are the same species but possess 2 SNVs in the middle part (286G, 298T for A11; 286A, 298T for A12), and our pipeline was able to assemble both into correct barcodes.

```
                286          298 300
A12_Hiseq  GCCTCCTTCTTTGACCCTACTTTTAGCTAGTAG
A12_sanger GCCTCCTTCTTTGACCCTACTTTTAGCTAGTAG
A11_Hiseq  ACCTCCTTCTTTGACCCTGCTTTTAGCTAGAAG
A11_sanger ACCTCCTTCTTTGACCCTGCTTTTAGCTAGAAG
```

Despite the success, we acknowledge the concerns about potential mis-assemblies, especially for subtle variations in the gap region. Therefore, we have modified our pipeline to also provide an additional sequence to accommodate possible alternative haplotypes and have the users notified by the program.

- "de brujin" is typo (page 7).

Response: Corrected.

Reviewer #2: General comments

In this study, the authors proposed an extremely efficient method for sequencing barcode DNA of identified specimens and for fattening out reference barcode sequence database.

This method may be very important and useful for barcoding, metabarcoding and mitometagenome skimming studies because the reference sequence database is crucial for bridging nucleotide sequences and taxonomic names and because taxonomic names are required for applying existing biological knowledges to barcoding, metabarcoding and mitometagenome skimming studies.

Thus, I strongly recommend to publish this study at Gigascience with several corrections of minor problems listed below.

The largest problem in this study is redistribution of USEARCH which is closed-source non-free software and redistribution is not allowed but included in the distributed file. Therefore, I recommend to replace USEARCH to VSEARCH which is free and open-source alternative of USEARCH or just exclude USEARCH from distribution.

Response: Many thanks for your kindly suggestion. We have replaced "USEARCH" with "VSEARCH" in our pipeline, and confirmed that VSEARCH provided identical results. The sentence "sequences were clustered using VSEARCH" was added in line 18 page 7.

The secondary problem is possibility of misassembly of very similar sequences. If misassembled sequences are registered to the reference sequence database, such sequences might cause misidentification of query sequences.

In order to avoid such possibilities, misassembled or misidentified sequences should be excluded from reference sequence database.

The proposed method assemble short-read Illumina sequences based on k-mer sequence matches and such misassembly was not observed in their real data, but it's still possible theoretically.

Thus, I recommend to add a function to warn users of a possibility of misassembly if same or similar scored assembly paths exist.

Such warning function can help users to detect problematic sequences.

Response: Many thanks again and we have modified our program so that it produces additional sequences which have similar or same scores comparing to their best alternative. The pipeline also provides an additional note file with a suffix of "note.txt" with notes alerting users about the possibilities. At last, we have also added explanation in the manual of our program.

Specific comments

P4L42 Add "of" to behind of "accuracy".

Response: Added.

P7L42 The authors wrote "much more sensitive" but did not write "than what?".

Response: We have modified this sentence to "the new pipeline is much more sensitive than Sanger in recovering amplicons at low quantity."

P7L60 Material -> Materials.

Response: Corrected.

P8L35 3uL of 10x reaction buffer was added but total reaction mixture was 25uL. Why?

Response: We do not completely understand the question. Perhaps the phrase "10x" is confusing. It refers to concentration of the buffer rather than volume. Therefore, in this

	<p>case, the volume of the reaction buffer is 3µl (of 10X concentration), and the total reaction mixture was 25µl.</p> <p>P8L40 I think this is not a "touchdown" PCR because the annealing temperature of first several cycles is lower than that of the following cycles. Response: "touchdown" has been removed.</p> <p>P8L60 Add "also" between "was" and "sequenced". Response: Added.</p> <p>P21L45 Add "illustration" between "Schematic" and "of". Response: Added.</p>
Additional Information:	
Question	Response
Are you submitting this manuscript to a special series or article collection?	No
Experimental design and statistics	Yes
<p>Full details of the experimental design and statistical methods used should be given in the Methods section, as detailed in our Minimum Standards Reporting Checklist. Information essential to interpreting the data presented should be made available in the figure legends.</p> <p>Have you included all the information requested in your manuscript?</p>	
Resources	Yes
<p>A description of all resources used, including antibodies, cell lines, animals and software tools, with enough information to allow them to be uniquely identified, should be included in the Methods section. Authors are strongly encouraged to cite Research Resource Identifiers (RRIDs) for antibodies, model organisms and tools, where possible.</p> <p>Have you included the information requested as detailed in our Minimum Standards Reporting Checklist?</p>	
Availability of data and materials	Yes
<p>All datasets and code on which the conclusions of the paper rely must be either included in your submission or deposited in publicly available repositories (where available and ethically appropriate), referencing such data using</p>	

a unique identifier in the references and in the "Availability of Data and Materials" section of your manuscript.

Have you have met the above requirement as detailed in our [Minimum Standards Reporting Checklist?](#)

1 **Filling reference gaps via assembling DNA barcodes using high-**
2 **throughput sequencing – moving toward barcoding the world**

3
4 4 Shanlin Liu^{1, 2, 4, #}, Chentao Yang^{2, #}, Chengran Zhou^{2, 5 #}, Xin Zhou^{1, 3, *}
5
6

7 6 1. Beijing Advanced Innovation Center for Food Nutrition and Human Health, College of
8 7 Plant Protection, China Agricultural University, Beijing 100193, People's Republic of
9 8 China

10 9 2. BGI-Shenzhen, Shenzhen, 518083, China

11 10 3. National Engineering Research Center for Fruit and Vegetable Processing, China
12 11 Agricultural University, Beijing 100193, People's Republic of China

13 12 4. Centre for GeoGenetics, Natural History Museum of Denmark, University of
14 13 Copenhagen, Copenhagen, Denmark

15 14 5. Key Laboratory of Bio-Resources and Eco-Environment, Ministry of Education, College
16 15 of Life Sciences, Sichuan University, Chengdu 610000, China.

17 16
18 17 #Equal contribution

19 18 *Corresponding to: Xin Zhou
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

1 **Abstract**

2
3
4 3 Over the past decade, biodiversity scientists have dedicated tremendous
5 efforts in constructing DNA reference barcodes for rapid species registration
6 and identification. Although analytical cost for standard DNA barcoding has
7 been significantly reduced since early 2,000, further dramatic reduction on
8 barcoding costs is unlikely because the Sanger sequencing is approaching its
9 limits in throughput and chemistry cost. Constraints in barcoding cost not only
10 led to unbalanced barcoding efforts around the globe, but also refrained High-
11 Throughput-Sequencing (HTS) based taxonomic identification from applying
12 binomial species names, which provide crucial linkages to biological
13 knowledge. We developed an Illumina-based pipeline, HIFI-Barcode, to
14 produce full-length COI barcodes from pooled PCR amplicons generated by
15 individual specimens. The new pipeline generated accurate barcode
16 sequences that were comparable to Sanger standards, even for different
17 haplotypes of the same species that were only a few nucleotides different
18 from each other. Additionally, the new pipeline was much more sensitive in
19 recovering amplicons at low quantity. The HIFI-Barcode pipeline successfully
20 recovered barcodes from over 78% of the PCR reactions that didn't show
21 clear bands on the electrophoresis gel. Moreover, sequencing results based
22 on the single molecular sequencing platform, Pacbio, confirmed the accuracy
23 of the HIFI-Barcode results. Altogether, the new pipeline can provide an
24 improved solution to produce full-length reference barcodes at about 1/10 of
25 the current cost, enabling construction of comprehensive barcode libraries for
26 local fauna, leading to a feasible direction for DNA barcoding global biomes.

1 **Background**

2 Over the past decade, biodiversity research has seen paradigm shifts in
3 methodology developments and applications [1], where standard DNA
4 sequences, e.g., DNA barcodes, are adopted for fast and accurate taxonomic
5 diagnoses, and High Throughput Sequencing (HTS) platforms are employed in
6 analysis of complex biological samples, including bulk samples [2, 3],
7 environmental DNA (eDNA, [4]), invertebrate-derived DNA (iDNA, [5, 6]) etc.
8 DNA barcode reference libraries have been constructed globally via synergistic
9 effort, resulting in well-curated, centralized barcode registration databases, e.g.,
10 the Barcode of Life Data systems [7], which has recently reached a milestone
11 for 5-million barcodes, covering ca. 0.26 million species (accessed in July 2017).
12 These DNA barcodes have been effectively facilitating species identification,
13 phylogenetic reconstruction [8], and understanding of interspecific interactions
14 and community structures [1].

15 Along with the rapid accumulation of global barcode references for various
16 taxon groups, significant effort has been made in digitalizing biomes, e.g.,
17 sequencing all taxa of particular lineages found in entire range of national parks
18 or islands [9]. Early efforts in barcoding biomes have employed standard
19 Sanger sequencing-based approaches to characterizing focal fauna [10-12].
20 Alternatively, boosted by HTS technologies, DNA metabarcoding and
21 mitochondrial metagenomics (mitochondrial genome skimming) have been
22 applied in investigations of local biodiversity and in evaluation of biological
23 managements [13-17]. These practices allow investigators to rapidly
24 understand species richness or even approximation for species evenness
25 and/or biomass for complex biological samples [4, 18]. A typical dilemma,
26 however, is the lack of local barcode references, from which HTS biodiversity
27 analysis could draw conclusions on species occurrences. This is primarily due
28 to unbalanced barcoding efforts around the globe, where regions in desperate
29 needs for biodiversity research are typically suffering from insufficient funding
30 for taxonomy work, especially for DNA based studies. Consequently, HTS-

1 based taxonomic registrations are often constraint to applying Molecular
2 Operational Units (MOTUs) instead of binomial species names, therefore
3 unable to associate existing biological and ecological knowledge to the
4 resultant diversity composition.

5 Admittedly, the analytical cost for standard DNA barcoding has been
6 significantly reduced since early 2000, a result from the development of
7 centralized and industrialized barcoding facilities and automated pipelines [1].
8 Currently, the average production cost for a reference barcode is ca. 10 USD,
9 excluding that for sample collection and handling. Further dramatic reduction
10 on barcoding costs is unlikely because Sanger sequencing technology is
11 approaching its limits in throughput and associated chemistry cost. It is
12 estimated that 100 million specimens would need to be sequenced to complete
13 the global barcode registration [1], which translates into a roughly 1-billion-
14 dollar budget for merely reference constructions. A similar challenge was seen
15 in the sequencing of the first human genome, where an initial budget of over 3
16 billion USD was estimated based on the application of Sanger sequencing [19].
17 Thanks to the advent of HTS technologies over the past decade, the current
18 cost for a human genome is within the range of just a few thousand USD, if not
19 less.

20 An early study using HTS in generating barcodes from single specimens
21 employed the Roche 454 platform [20], which was rapidly phased out due to
22 limited throughput capacity (hence high chemistry cost). Illumina platforms (e.g.,
23 Hiseq and Miseq) have been primarily applied in recent practices [21]; but these
24 are constraint by relatively short read lengths (100-300 bps). Even with the most
25 recent Miseq model at 300bp paired-end (PE) sequencing, full-length barcodes
26 (e.g., ~700 bps for COI including primers) are beyond the sequencing range.
27 Therefore, existing pipelines are forced to produce a fragment of the standard
28 barcodes (e.g., 313bp,[22]) or to apply 2 rounds of PCR amplifications, each
29 targeting on a proportion of the full barcodes [21]. Obviously, full-length
30 barcodes are desired for constructing barcode references and extra

1 amplification procedure should be avoided when possible for cost control and
2 simplification of pipelines. In particular, efficient primers might be difficult to
3 identify in the mid-COI barcode region across taxon groups. Alternatively, short
4 HTS reads can be assembled into much longer scaffolds, which is a standard
5 practice in *de novo* genome or transcriptome assembling. In fact, a specific
6 assembly algorithm, SOAPBarcode, has been developed for recovering full-
7 length barcodes from pooled arthropod samples [23].

8 Here, we introduce a more straightforward and cost-efficient HTS pipeline that
9 generates full-length reference barcodes - HIFI-Barcode (Fig. 1). Briefly,
10 individual genomic DNA was extracted separately and amplified on a 96-well
11 plate using 96 sets of uniquely tagged primers. Amplicons were then pooled
12 and sequenced on an Illumina Hiseq 4000 platform at 150 paired-end (PE).

13 Mixed HTS reads were assembled using a customized bioinformatics pipeline
14 to obtain barcode sequence for each individual. Compared to aforementioned
15 studies [21, 22], our method can deliver standard full-length barcodes via a
16 single PCR reaction and the sequencing is carried out on HiSeq platforms, the
17 most cost-effective HTS platform currently available. Using Sanger barcodes
18 as the gold standard, the new pipeline can generate accurate individual
19 barcode sequences, even for haplotypes of the same species that are only a
20 few nucleotides different from each other. Additionally, the new pipeline is much
21 more sensitive than Sanger in recovering amplicons at low quantity. Over 78%
22 (25/32) of the “failed” PCR amplicons (those without clear bands on an
23 electrophoresis gel) were successfully recovered at high-quality using the new
24 pipeline. In addition, the single-molecule sequencing platform, Pacbio, has also
25 been adopted in our study to evaluate the accuracy of the HIFI-Barcode method.
26 Altogether, the new pipeline can provide an alternative solution to produce full-
27 length reference barcodes at about 1/10 of the current cost, enabling larger-
28 scale biodiversity barcoding initiatives, especially for areas where DNA
29 references are scarce.

1 **Materials and methods**

2 1. DNA preparation

3 Insect specimens were collected in Laohegou Natural Reserve, Sichuan
4 Province, China. Genomic DNA was extracted in an independent study using
5 the Glass Fiber Plate method following manufacturer's protocol [24]. Two 96-
6 well plates were prepared for the current work: one plate containing 96 high-
7 quality lepidopteran DNA (showing clear band of standard DNA barcode
8 amplicon on an electrophoresis gel) is used to evaluate the accuracy of our
9 HTS method using Sanger barcodes as the gold standard; a second plate
10 containing 95 randomly selected DNA (mostly dipterans) regardless of quality
11 and PCR yields plus a blank control is prepared to examine the success rate
12 of our HTS method compared with the classic Sanger approach.

14 2. DNA amplification and sequencing

15 Ninety-six pairs of different tags were added to both ends of a common COI
16 barcode primer set (LCO1490 and HCO2198, Supplemental Table S1) [25],
17 with each tag containing 5 bps allowing for ≥ 3 bp differences from each other.
18 Each PCR reaction contained 1 μ l of DNA template, 16.2 μ l of molecular biology
19 grade water, 3 μ l of 10X reaction buffer (Mg^{2+} plus), 2.5 μ l of dNTPs mix (10
20 mM), 1 μ l of forward and reverse primers (10 mM), and 0.3 μ l of TaKaRa Ex
21 Taq polymerase (5 U/ μ l). The amplification program included a thermocycling
22 profile of 94°C for 1 min, 5 cycles of 94°C for 30 sec, 45°C for 40 sec, and an
23 extension at 72°C for 1 min, followed by 35 cycles of 94°C for 30 sec, 51°C for
24 40 sec, and 72°C for 1 min, with a final extension at 72°C for 10 min, and finally
25 holding at 12°C. All amplicons were visualized on a 1.2% 96 Agarose E-gel
26 (Biowest Agarose). All PCR products from each plate were pooled using 1 μ l
27 per sample resulting in two 96 μ l mixtures, which were sent to BGI and
28 sequenced using a Hiseq 4000. PCR amplicons were fragmented to construct
29 library of an insert-size of 250 bp and sequenced with a strategy of 150 PE. A
30 second set of PCR mixture of the 2nd plate (576 μ l, 6 μ l per sample) was also

1 sequenced using PacBio RS II at NextOmics.

3. HIFI-Barcode assembly

4 *Data filtering:* Reads of bad-quality were removed from raw data: 1) reads with
5 adapter contamination (≥ 15 bps alignment length and ≤ 3 mismatches); 2) reads
6 with >10 Ns; 3) reads with >50 bps of low quality (Phred quality score = 2, ASCII
7 35 "B", Illumina 1.8+ Phred+33).

8
9 *Read assignment:* Firstly, reads containing 5' and 3' ends of each individual
10 were identified based on their unique 5-mer tags and corresponding primer
11 sequences using in house Perl scripts. Then, for each individual, identical reads
12 were clustered to obtain unique 5' and 3' sequences. Each individual may
13 contain multiple unique terminal sequences at varied abundances due to
14 haplotype heterogeneity (mitochondrial heteroplasmy) or artefacts (PCR or
15 sequencing errors). Next, the most abundant unique sequence was chosen for
16 the following overlapping and assembly procedures. In addition, if the next most
17 abundant unique sequence had an abundance $\geq 1/10$ of that of the most
18 abundant unique sequence at $<98\%$ similarity (sequences were clustered using
19 VSEARCH [26]), it was also retained to confirm identities, e.g., parasites,
20 *Wolbachia*, gut contents that were co-amplified in PCR. After that,
21 corresponding pairs of the afore-chosen reads were identified according to their
22 titles and then paired-end reads were overlapped using COAP [27] with an
23 identity cutoff of 95%. Overlapped reads could vary in sequence length due to
24 insert-size fluctuation during ultrasonic shearing. Thus, consensus 5' and 3'
25 sequences of each individual were achieved using in house Perl scripts where
26 ends with read coverage < 5 were trimmed off (Fig. 2).

27
28 *Gap filling:* Algorithm adopted from SOAPBarcode [23] (Supplemental Fig. S1)
29 was applied to fill the gaps between 5' and 3' terminal scaffolds of each
30 individual to complete the full-length barcodes. Briefly, for each individual, the

1 5' end was defined as the start point, and the 3' end as the end point. Then, the
2 kmer set from de Bruijn graph was walked step by step from the start point to
3 the end point to find potential assembly paths. Several strategies were applied
4 to ensure correct paths: 1) kmers of abundance < 10% of the average kmer
5 abundance before path bifurcation were removed; 2) if there was more than
6 one out degree remaining after step one, common reads were counted between
7 different out degrees and the kmer located before the last bifurcation, and the
8 out degrees of common reads < 10% of the average abundance were removed;
9 3) paths expanding beyond the pre-set length (standard COI barcode length
10 plus primers) without an end point were removed.

11 12 5. Data filtering and read assignment for Pacbio

13 Pacbio SmrtAnalysis pipeline (<https://github.com/PacificBiosciences>) was
14 adopted to extract 28,770 circular consensus sequencing (CCSs) from 1.1G
15 raw data. Then, CCSs of ≥ 15 passes were chosen for next steps: 1) 22,075
16 CCSs were demultiplexed by their corresponding indices using an in-house Perl
17 script, allowing a maximum of 1bp deletion at the 5' end of forward index or the
18 3' end of reverse index. 2) for each sample, sequences with a length range out
19 of 658 ± 6 bp were removed and the remaining unique sequences were sorted
20 by pass numbers and identical sequences were clustered together; 3) unique
21 sequence of the most abundant cluster was retained as the correct barcode
22 sequence for each sample.

23 24 6. Comparisons between HTS, Sanger barcodes and Pacbio clusters

25 Barcode sequences obtained by Sanger, HIFI-Barcode method and Pacbio
26 were subject to phylogenetic tree constructions using MEGA7 (Neighbor-joining
27 and 1,000 bootstrap) and iTOL [28]. BWA [29] was applied to align raw reads
28 to assembled HTS barcodes to examine discrepancies between HTS and
29 Sanger sequences.

1 Results

2 A total of 4,824,443 and 4,439,345 PE reads for the 1st and 2nd plate were
3 obtained after data filtering, respectively, using Hiseq 4000.

4 For the 1st plate, a total of 1,910,616 (39.60%) reads were assigned to their
5 corresponding samples as either 5' or 3' end, and 1,898,372 (39.34%) as reads
6 belonging to intermediate regions, while 1,015,455 (21.05%) reads were
7 identified as primer dimers or short PCR chimeras. The abundance of end
8 reads for each sample varies significantly, ranging from 2,444 to 64,705. After
9 clustering at 100%-similarity for the 5' and 3' end reads, most samples (61 out
10 of 96) obtained single unique reads after read assignment. The 2nd plate
11 possessed similar read distribution and details of both plates were summarized
12 in Table 1.

13 One cell of Pacbio data containing 28,770 circular consensus sequencing (CCS)
14 from 1,201,158 raw reads were generated for the 2nd plate. CCS reads had an
15 average pass number of 26.5 and were assigned to 82 samples after
16 demultiplexing. (Table 1). Note that a single Pacbio sequencing read can reach
17 as long as 40 kb. Therefore, short CCS read of high quality can be sequenced
18 dozens of times, which in turn effectively corrects sequence errors associated
19 with the platform [30].

20 21 *Accuracy and Efficiency:*

22 Sanger barcodes were obtained from all 96 lepidopteran samples of the 1st
23 plate (Fig. 3A), including 91 haplotypes and 85 OTUs using a similarity
24 threshold of $\geq 98\%$. The HIFI-Barcode assemblies were successful for all 96
25 samples and showed high accuracy compared to Sanger sequences. Even
26 identical or highly similar barcodes from individuals of the same species were
27 correctly assembled, e.g., A2 versus F7, B1 versus E1, and C7 versus G4 (Fig.
28 3B and 3C). A total of 43 ambiguous sites (out of 63,168 bps) found in Sanger
29 barcodes were identified to a specific nucleotide in HIFI barcodes (e.g., Fig. 3D,
30 Fig. 4B). Only 9 HIFI barcodes showed a single nucleotide difference from the

1 corresponding Sanger sequences, which could reflect ambiguous base-calling
2 in Sanger sequencing or genuine heteroplasmy in the examined individual. At
3 least two of the discrepancies were proven to be heteroplasmy via mapping raw
4 reads against discrepant sites (Fig. 4A).

5 In the 2nd plate, samples were randomly selected regardless of their DNA quality
6 and PCR success rates. Sixty-three PCR reactions showed clear bands on the
7 electrophoresis gel (Supplemental Table S2), of which 62 resulted in Sanger
8 barcodes. The HIFI-Barcode pipeline successfully produced full-length HTS
9 sequences for all 62 corresponding Sanger barcodes at high accuracy (56 at
10 100% match, 5 with 1 mismatch, and 1 with 3 mismatches, Supplemental Fig.
11 S2). In addition, HIFI barcodes were successfully generated from 25 out of the
12 32 PCR amplicons that had no clear bands (Supplemental Table S2), increasing
13 the overall success rate from 66.32% to 92.63%, for the Sanger and HIFI-
14 Barcode methods, respectively (Fig. 5). To further evaluate the accuracy of the
15 newly developed HIFI-Barcode pipeline especially for those where PCR
16 reactions failed, we also sequenced pooled PCR amplicons using Pacbio. The
17 CCSs used in our study had pass numbers > 15, which meant the same
18 molecule was sequenced repeatedly for more than 15 times. Thus the
19 consensus nucleotides for each sequence were corrected from sequencing
20 errors associated with the platform (ca. 10% on average). The overall success
21 rates for Pacbio was 86.32%. Of the 25 HIFI barcodes where Sanger failed, 18
22 Pacbio barcodes were obtained. Among these, 10 were identical to the
23 corresponding HIFI barcodes; 3 had 1 or 2 sites matched with one of the two
24 heterozygous alleles from HIFI barcodes; and 5 showed errors in amino acid
25 translation (e.g., stop codon) possibly due to sequencing errors in Pacbio
26 (Supplemental Table S3 and Supplemental file S1).

27 28 *Non-target sequences detected by HIFI-Barcode*

29 During the HIFI-Barcode assembly procedure, terminus sequences with $\geq 1/10$
30 abundance of that of the most abundant scaffolds at <98% similarity were

1 retained for assembly and identity check. This analysis allowed detection of 18
2 non-target sequences co-amplified from the 2 plates (Supplemental Table S4),
3 in addition to COI barcodes. Cross examinations against both NCBI and
4 barcode sequences from the focal plates suggested origins including
5 *Wolbachia* (2), fungus (1), cross-contamination from adjacent wells (7), as well
6 as potential PCR errors and pseudo-genes (8). The presence of non-target
7 PCR products from the 2nd plate were further confirmed by Pacbio sequencing
8 at >99% identity, therefore ruling out the likelihood of assembly errors in the
9 HIFI-Barcode pipeline, suggesting they are co-amplified numts present in PCR
10 products. These low-quantity sequences are likely common in regular PCR-
11 based pipelines and detectable by HTS-based approaches. But they can be
12 easily filtered out from genuine COI barcodes following the pipeline described
13 in this study.

14 **Discussion**

15 It is widely acknowledged that we have been undergoing unprecedented global
16 biodiversity loss [31]. DNA-based approaches, e.g., DNA barcoding, DNA
17 metabarcoding, mitochondrial metagenomics (mitochondrial genome
18 skimming), have demonstrated efficacy in accelerating biodiversity inventories
19 of large geographical ranges. These standardized and largely automated
20 procedures will provide pivotal information to understand how biodiversity loss
21 is characterized and how to desist from it. New methodologies enable rapid
22 collection of biodiversity and ecology data at large scale over space and time,
23 which in-turn benefits policy-makers at varied management levels and research
24 groups [32].

25 Interpreting molecular results using existing knowledge on biology, ecology and
26 evolution would require a linkage between DNA references and Linnaeus
27 names, which is one of the fundamental roles of DNA barcoding initiatives. The
28 construction of comprehensive barcode references is still, to a large extent,
29 expensive and sometimes prohibitive. This is particularly true for studies

1 targeting on a wide range of taxa from a large area of natural habitat. Although
2 the most represented DNA barcode database (BOLD) now hosts barcodes for
3 0.26 million species, accounting for ca. 1/4 of described species, chances of
4 encountering a novel barcode are still very high, especially for many biodiversity
5 hotspots. Even if an ecological study focuses on just a small proportion of the
6 focal diversity, it is not uncommon that hundreds to thousands of species would
7 need to be barcoded to draw meaningful conclusions. In addition, multiple
8 individuals of the same species (ideally from distinct populations) would need
9 to be sequenced to reflect intraspecific genetic diversities. There is no
10 consensus on the ideal number of conspecific individuals to be sequenced, but
11 in practice an average of 10 is often followed, while some study recommends
12 20 [33], if not a lot more. Therefore, roughly tens of thousands of individuals,
13 requiring hundreds of thousands of USD, are expected to suffice a regular
14 ecology study, just for the molecular analysis (for a recent example, please see
15 [12]). While the HTS-based approaches have shown promising power in
16 analyzing complex sample mixtures at much reduced unit cost [2-4, 14], one
17 would still need to establish DNA barcode references to be able to go beyond
18 OTU-based interpretation.

19 The HIFI-Barcode method, as the results showed, offers a novel route to
20 produce mass volumes of reliable barcode sequences at significantly reduced
21 cost. The main costs of the HIFI-Barcode pipeline include consumable
22 chemistries, library construction, high-throughput sequencing, and informatics.
23 Despite the increased one-time cost in ordering multiple unique sets of primers,
24 the cost on primers per unit reaction is negligible. Following our protocols, the
25 average cost for a HIFI barcode is around 1 USD, as opposed to 10-20 USD in
26 the standard Sanger approach. Further saving on the production cost is
27 achieved by increased success rates, especially for amplicons with low quantity.
28 In our test, ca. 1/3 of the 2nd plate would have been re-amplified in standard
29 barcoding protocols, using a different set of primers, followed by gel
30 examination, positive picking, PCR purification, and Sanger-sequencing.

1 By complementing the barcode reference library at <1/10 of the current cost,
2 the new approach also reinforces rapid constructions of organelle genomes,
3 e.g., mitochondria and chloroplasts. A number of pilot studies have
4 demonstrated that full mitochondrial genomes can provide elevated power in
5 bulk sample analysis [18, 34]. New approaches to assembling full mito-
6 genomes or the majority of the coding genes have been developed for shotgun
7 sequencing of individual specimens [35], pooled taxa [18, 34], and
8 transcriptomes [35]. In particular, mito-genome assembly through direct
9 shotgun sequencing of mixed taxa can significantly reduce the library
10 construction cost for HTS. Bait sequences, which regularly includes standard
11 COI barcodes, are important for assigning mixed mitochondrial scaffolds to a
12 specific taxon. This is critical especially if the phylogenetic signal of the
13 scaffolds alone is not sufficient to attribute assemblies to species, e.g., when
14 multiple closely related species are pooled. In fact, having multiple bait
15 sequences per species will significantly remove bioinformatics challenge during
16 the assembly procedure [36], which now becomes financially feasible with the
17 help of the HIFI-Barcode pipeline.

18
19 Several aspects can be further improved for our method: 1) Multiple barcode
20 markers (e.g., COI, CYTB, 12S, etc.) can be pooled in a single shotgun
21 sequencing effort without increasing tag complexity, which will again alleviate
22 analytical cost. 2) The pooled PCR amplicons were subject to library
23 construction directly in the present study. The proportion of primer dimers and
24 short PCR chimera reached as high as ca. 21% in our raw reads, which can be
25 easily removed using size-preference magnetic beads. 3) Addition of inosine to
26 the 3' terminus of the primer may increase its universality and will further elevate
27 the successful rate and efficiency. 4) Longer tags allowing for pooling more
28 individuals (e. g. 384-well plate) can further increase the throughput capacity.

29 In summary, the HIFI-Barcode method provides a HTS-based approach with
30 improved economic efficiency, which allows investigators to produce standard

1 full-length barcodes at ca. 1/10 of the current cost. The new protocol not only
2 generates barcode sequences of high quality that are comparable to Sanger-
3 barcodes, but also increases overall sequencing success rates by detecting
4 PCR amplicons in minute quantities. This new method enables construction of
5 comprehensive barcode libraries for local fauna, leading to a feasible direction
6 for DNA barcoding global biomes.

8 **Availability of source code and requirements**

- 9 • Project name: HIFI–Barcode project
- 10 • Operating system(s): Unix, Linux
- 11 • Programming language: PERL
- 12 • Other requirements: GCC version $\geq 4.4.5$
- 13 • License: GNU General Public License version 3.0 (GPLv3)
- 14 • Any restrictions to use by non-academics: none

16 **Availability of Supporting Data**

17 Availability of supporting data and materials code and data are available in the
18 GigaScience GigaDB repository (XXX). Source code also can be found in
19 <https://github.com/comery/HIFI-barcode-hiseq> and
20 <https://github.com/comery/HIFI-barcode-pacbio>.

22 **Abbreviations**

23 BOLD - the Barcode of Life Data systems; CCSs – circular consensus
24 sequencing; eDNA – environmental DNA; HTS – High Throughput Sequencing;
25 GB – Gigabase; iDNA – invertebrate-derived DNA; MOTUs – Molecular
26 Operational Units; PCR - Polymerase Chain Reaction; PE – Paired End.

28 **Acknowledgements**

29 We thank Dr. Ding Yang from China Agricultural University for contributing
30 samples and Yongfei Du from BGI-Shenzhen for his help on coordinating the

1 HiSeq sequencing of our 2nd plate. We would like to thank Min Tang, Guanliang
2 Meng for assistant in figure preparation. This work is supported by the China
3 National GeneBank, BGI and by China Agricultural University through the
4 Chinese Universities Scientific Fund, 2017QC114 to XZ.

6 **Author contributions:**

7 X.Z. and S.L. designed the study; S.L. coordinated the project and led the
8 analyses; C.Z. and C.Y. led the bench work and contributed to the analyses;
9 S.L., C.Z., and C.Y. formulated the early drafts, and X.Z. revised the manuscript.

11 **Competing interests**

12 The authors declare that they have no competing interests.

14 **References**

- 16 1. Hebert PD, Hollingsworth PM, Hajibabaei M: **From writing to reading the**
17 **encyclopedia of life.** *Phil Trans R Soc B* 2016, **371**(1702):20150321.
- 18 2. Yu DW, Ji YQ, Emerson BC, Wang XY, Ye CX, Yang CY, Ding ZL: **Biodiversity Soup:**
19 **Metabarcoding of arthropods for rapid biodiversity assessment and**
20 **biomonitoring.** *Methods in Ecology and Evolution* 2012, **3**(4):613-623.
- 21 3. Zhou X, Li Y, Liu S, Yang Q, Su X, Zhou L, Tang M, Fu R, Li J, Huang Q: **Ultra-deep**
22 **sequencing enables high-fidelity recovery of biodiversity for bulk arthropod**
23 **samples without PCR amplification.** *GigaScience* 2013, **2**(1):4.
- 24 4. Taberlet P, Coissac E, Hajibabaei M, Rieseberg LH: **Environmental DNA.** *Molecular*
25 *Ecology* 2012, **21**:1789–1793.
- 26 5. Calvignac - Spencer S, Merkel K, Kutzner N, Kühl H, Boesch C, Kappeler PM, Metzger
27 S, Schubert G, Leendertz FH: **Carion fly - derived DNA as a tool for**
28 **comprehensive and cost - effective assessment of mammalian biodiversity.**
29 *Molecular Ecology* 2013, **22**(4):915-924.
- 30 6. Schnell IB, Thomsen PF, Wilkinson N, Rasmussen M, Jensen LR, Willerslev E,
31 Bertelsen MF, Gilbert MTP: **Screening mammal biodiversity using DNA from**
32 **leeches.** *Current biology* 2012, **22**(8):R262-R263.
- 33 7. Ratnasingham S, Hebert PDN: **BOLD: The Barcode of Life Data System ([http://www.](http://www.barcodinglife.org)**
34 **barcodinglife.org).** *Molecular ecology notes* 2007, **7**(3):355-364.
- 35 8. Zhou X, Frandsen PB, Holzenthal RW, Beet CR, Bennett KR, Blahnik RJ, Bonada N,
36 Cartwright D, Chuluunbat S, Cocks GV: **The Trichoptera barcode initiative: a**
37 **strategy for generating a species-level Tree of Life.** *Phil Trans R Soc B* 2016,

1 371(1702):20160025.

- 2 9. Miller SE, Hausmann A, Hallwachs W, Janzen DH: **Advancing taxonomy and**
3 **bioinventories with DNA barcodes.** *Phil Trans R Soc B* 2016, **371**(1702):20150339.
- 4 10. Zhou X, Robinson JL, Geraci CJ, Parker CR, Flint Jr OS, Etnier DA, Ruitter D, DeWalt
5 RE, Jacobus LM, Hebert PD: **Accelerated construction of a regional DNA-barcode**
6 **reference library: caddisflies (Trichoptera) in the Great Smoky Mountains**
7 **National Park.** *Journal of the North American Benthological Society* 2011, **30**(1):131-
8 162.
- 9 11. Zhou X, Adamowicz SJ, Jacobus LM, DeWalt RE, Hebert PD: **Towards a**
10 **comprehensive barcode library for arctic life-Ephemeroptera, Plecoptera, and**
11 **Trichoptera of Churchill, Manitoba, Canada.** *Frontiers in zoology* 2009, **6**(1):1.
- 12 12. Telfer AC, Young MR, Quinn J, Perez K, Sobel CN, Sones JE, Levesque-Beaudin V,
13 Derbyshire R, Fernandez-Triana J, Rougerie R: **Biodiversity inventories in high gear:**
14 **DNA barcoding facilitates a rapid biotic survey of a temperate nature reserve.**
15 *Biodiversity data journal* 2015(3).
- 16 13. Valentini A, Taberlet P, Miaud C, Civade R, Herder J, Thomsen PF, Bellemain E,
17 Besnard A, Coissac E, Boyer F: **Next - generation monitoring of aquatic**
18 **biodiversity using environmental DNA metabarcoding.** *Molecular Ecology* 2016.
- 19 14. Hajibabaei M, Shokralla S, Zhou X, Singer GAC, Baird DJ: **Environmental barcoding:**
20 **a next-generation sequencing approach for biomonitoring applications using**
21 **river benthos.** *PloS one* 2011, **6**(4):e17497.
- 22 15. Coissac E, Riaz T, Puillandre N: **Bioinformatic challenges for DNA metabarcoding**
23 **of plants and animals.** *Molecular Ecology* 2012, **21**:1834–1847.
- 24 16. Bienert F, Dedanieli S, Miquel C, Coissac E, Poillot C, Brun JJ, Taberlet P: **Tracking**
25 **earthworm communities from soil DNA.** *Molecular Ecology* 2012, **21**(8):2017-2030.
- 26 17. Baird DJ, Hajibabaei M: **Biomonitoring 2.0: a new paradigm in ecosystem**
27 **assessment made possible by next-generation DNA sequencing..** *Molecular*
28 *Ecology* 2012, **21**(8):2039-2044.
- 29 18. Gillett CP, Crampton-Platt A, Timmermans MJ, Jordal B, Emerson BC, Vogler AP: **Bulk**
30 **de novo mitogenome assembly from pooled total DNA elucidates the phylogeny**
31 **of weevils (Coleoptera: Curculionoidea).** *Molecular biology and evolution*
32 2014:msu154.
- 33 19. Krulwich R, LANDER E: **Cracking the Code of Life.** In.: *Public Broadcasting Service;*
34 2001.
- 35 20. Shokralla S, Gibson JF, Nikbakht H, Janzen DH, Hallwachs W, Hajibabaei M: **Next -**
36 **generation DNA barcoding: using next - generation sequencing to enhance and**
37 **accelerate DNA barcode capture from single specimens.** *Molecular Ecology*
38 *Resources* 2014, **14**(5):892-901.
- 39 21. Shokralla S, Porter TM, Gibson JF, Dobosz R, Janzen DH, Hallwachs W, Golding GB,
40 Hajibabaei M: **Massively parallel multiplex DNA sequencing for specimen**
41 **identification using an Illumina MiSeq platform.** *Scientific reports* 2015, **5**.
- 42 22. Meier R, Wong W, Srivathsan A, Foo M: **\$1 DNA barcodes for reconstructing**
43 **complex phenomes and finding rare species in specimen - rich samples.**
44 *Cladistics* 2016, **32**(1):100-110.

- 1 23. Liu S, Li Y, Lu J, Su X, Tang M, Zhang R, Zhou L, Zhou C, Yang Q, Ji Y: **SOAPBarcode: revealing arthropod biodiversity through assembly of Illumina shotgun sequences of PCR amplicons.** *Methods in Ecology and Evolution* 2013, **4**(12):1142-1150.
- 2
3
4
5 24. Ivanova NV, Dewaard JR, Hebert PD: **An inexpensive, automation - friendly protocol for recovering high - quality DNA.** *Molecular Ecology Resources* 2006, **6**(4):998-1002.
- 6
7
8
9 25. Vrijenhoek R: **DNA primers for amplification of mitochondrial cytochrome c oxidase subunit I from diverse metazoan invertebrates.** *Molecular marine biology and biotechnology* 1994, **3**(5):294-299.
- 10
11
12
13 26. Rognes T, Flouri T, Nichols B, Quince C, Mahé F: **VSEARCH: a versatile open source tool for metagenomics.** *PeerJ* 2016, **4**:e2584.
- 14
15
16 27. Liu B, Yuan J, Yiu SM, Li Z, Xie Y, Chen Y, Shi Y, Zhang H, Li Y, Lam TW: **COPE: an accurate k-mer-based pair-end reads connection tool to facilitate genome assembly.** *Bioinformatics* 2012, **28**(22):2870-2874.
- 17
18
19
20 28. Letunic I, Bork P: **Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees.** *Nucleic Acids Research* 2016:gkw290.
- 21
22
23
24 29. Li H, Durbin R: **Fast and accurate short read alignment with Burrows–Wheeler transform.** *Bioinformatics* 2009, **25**(14):1754-1760.
- 25
26
27 30. Eid J, Fehr A, Gray J, Luong K, Lyle J, Otto G, Peluso P, Rank D, Baybayan P, Bettman B: **Real-time DNA sequencing from single polymerase molecules.** *Science* 2009, **323**(5910):133-138.
- 28
29
30
31 31. Pimm SL, Jenkins CN, Abell R, Brooks TM, Gittleman JL, Joppa LN, Raven PH, Roberts CM, Sexton JO: **The biodiversity of species and their rates of extinction, distribution, and protection.** *Science* 2014, **344**(6187):1246752.
- 32
33
34
35 32. Díaz S, Demissew S, Carabias J, Joly C, Lonsdale M, Ash N, Larigauderie A, Adhikari JR, Arico S, Báldi A: **The IPBES Conceptual Framework—connecting nature and people.** *Current Opinion in Environmental Sustainability* 2015, **14**:1-16.
- 36
37
38
39 33. Luo A, Lan H, Ling C, Zhang A, Shi L, Ho SY, Zhu C: **A simulation study of sample size for DNA barcoding.** *Ecology and Evolution* 2015, **5**(24):5869-5879.
- 40
41
42
43 34. Tang M, Tan M, Meng G, Yang S, Su X, Liu S, Song W, Li Y, Wu Q, Zhang A: **Multiplex sequencing of pooled mitochondrial genomes—a crucial step toward biodiversity analysis using mito-metagenomics.** *Nucleic Acids Research* 2014, **42**(22):e166-e166.
- 44
45
46
47 35. Tan M, Zhang R, Xiang C, Zhou X: **The complete mitochondrial genome of the pumpkin fruit fly, *Bactrocera tau* (Diptera: Tephritidae).** *Mitochondrial DNA Part A* 2016, **27**(4):2502-2503.
- 48
49
50
51 36. Timmermans MJ, Dodsworth S, Culverwell C, Bocak L, Ahrens D, Littlewood DT, Pons J, Vogler AP: **Why barcode? High-throughput multiplex sequencing of mitochondrial genomes for molecular systematics.** *Nucleic acids research* 2010, **38**(21):e197-e197.
- 52
53
54
55
56
57
58
59
60
61
62
63
64
65

16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

Table 1. Read distribution of both Illumina and Pacbio platforms

	Raw read	Clean read	5' and 3' read	Read in-between	Recovered Indices	Sample size ¹	Single unique ²	Full-length barcodes
Hiseq 1	8,567,336	4,824,443	1,910,616	1,898,372	96	39,805 (64,705; 2,444)	61	96
Hiseq 2	11,531,498	4,439,345	1,306,054	2,676,915	96	27,210 (101,512; 279)	45	88
Pacbio 2*	1,201,158	28,770	26.4	17,102	82	208 (1,696; 1)	NA	82
		Total number 3	Average pass 3	Assigned ³				

Note: * number 1 and 2 in this column represent plate ID; 1. Read number possessed by samples in format as: average (max; min); 2. Number of clusters that left only 1 single representative candidate after read assignment filtering; 3. Statistics of Circular Consensus Sequence (CCS)

Figure legends

Figure 1. Schematic illustration of HIFI-Barcode pipeline.

Figure 2. HIFI-Barcode assembly pipeline.

Figure 3. Comparison between HIFI-Barcode and Sanger reference. 3A, success rates of the 1st plate. For all 96 samples, both Sanger (left semicircle) and HIFI-Barcode (right semicircle) are successful in producing a full-length COI barcode. Samples with red out lining are marked on the phylograms; 3B, phylogenetic tree of all HIFI barcodes and Sanger references; 3C, close-up view of representative individuals; 3D, degenerate sites of Sanger references were recuperated by HIFI barcodes.

16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

Figure 4. Discrepancies between Sanger and HIFI barcodes in the 1st plate. Entropy weight was calculated based on the strength of read depth by aligning Illumina raw reads onto assembled HIFI-barcodes, showing potential heteroplasmy (4A) and differences between ambiguous Sanger base-calling and specific nucleotide identify in HIFI barcodes (4B).

Figure 5. Success rates of the 2nd plate. For each sample, the upper, left and right pies represent PCR, HIFI-Barcode and Pacbio, respectively. Gray represents failure and the others represent success.

Additional Files

Supplemental Figure S1. Algorithm described in SOAPBarcode pipeline.

Supplemental Figure S2. Phylogenetic tree of samples sharing Sanger references, HIFI barcodes and Pacbio barcodes.

Supplemental Table S1. Indexed Primer sequences.

Supplemental Table S2. PCR electrophoresis results.

Supplemental Table S3. Comparison of 18 Pacbio barcodes and HIFI-barcodes

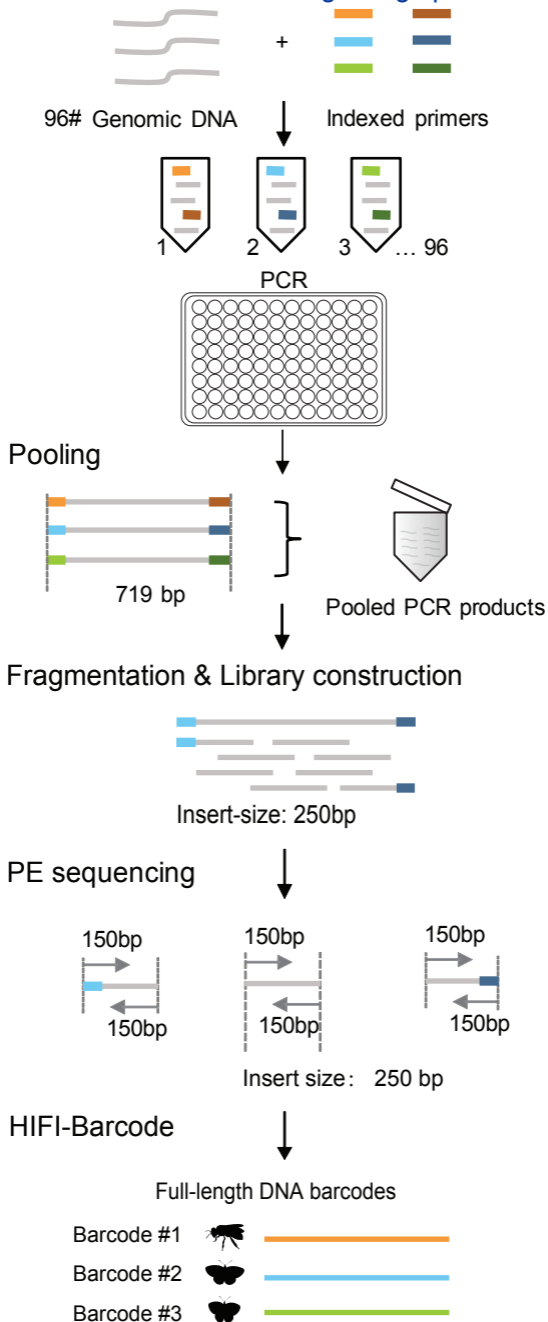
Supplemental Table S4. Non-target sequences detected by HIFI-Barcode

Supplemental File S1. Results of HIFI-barcode

Figure 1

Indexed PCR

[Click here to download Figure fig1.pdf](#)



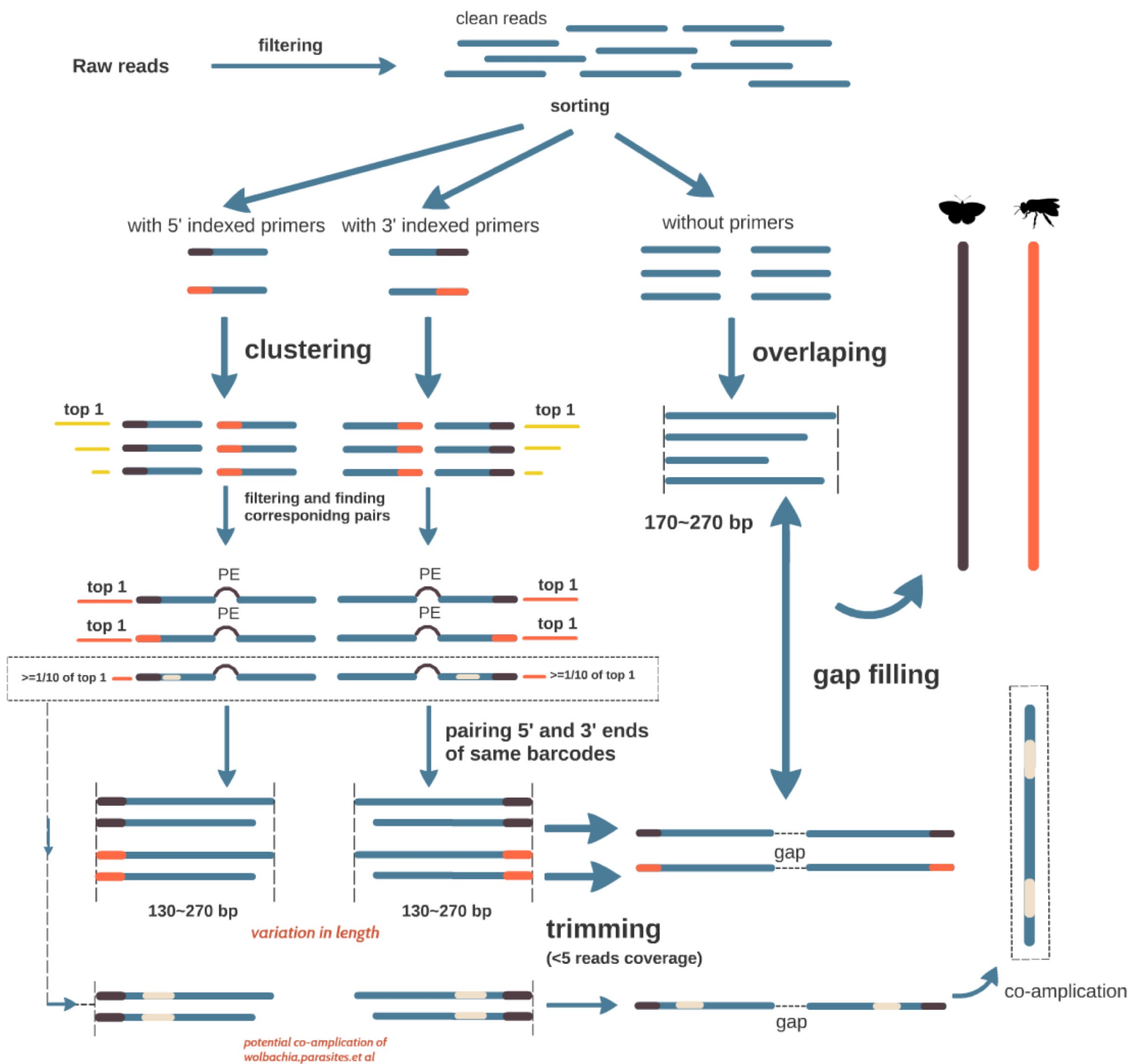
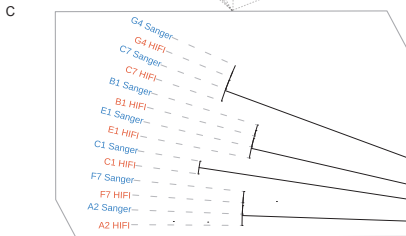
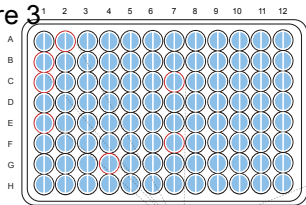


Figure 3



B [Click here to download Figure fig3.pdf](#)



D

<i>Neptis sappho</i> :	B6_Sanger	tatwttattaacagatcggaatmwwaatacyycatttttygaycctgctgraggwgg
	B6_HIFI	tataTTattaacagatcggaatCTtaatacTtcattttttgatCctgctggaggagg
<i>Aporia genestieri</i> :	B1_Sanger	acarrtataccccccctcttttctctataatatkGctcayggaggctcctctgtagattta
	B1_HIFI	acagTataccccccctcttttctctataatattGctcatggaggctcctctgtagattta

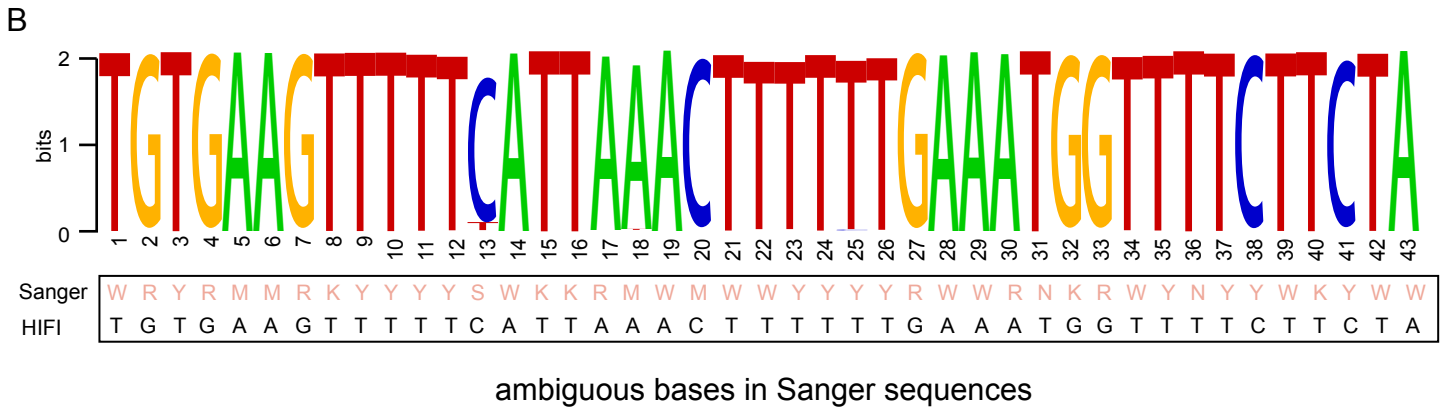
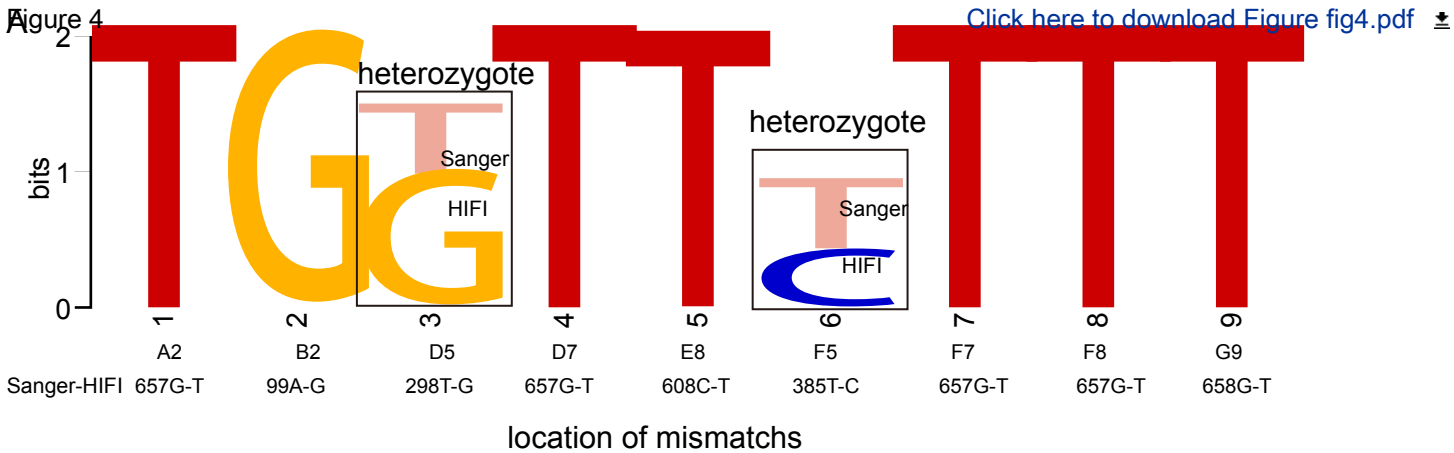


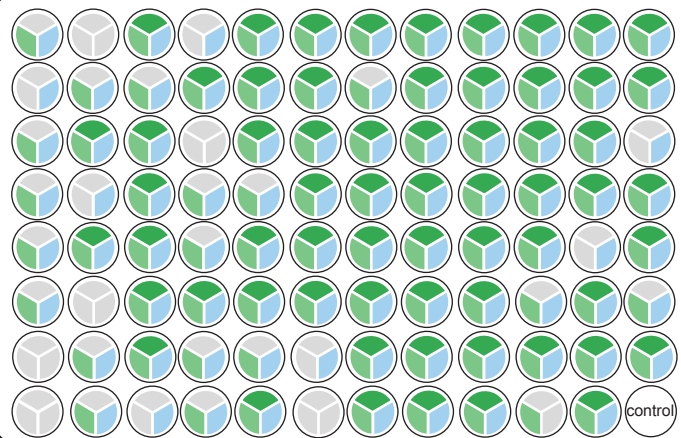
Figure 5

1 2 3 4 5 6 7 8 9 10 11 12

[Click here to download Figure fig5.pdf](#)



A
B
C
D
E
F
G
H









Click here to access/download
Supplementary Material
supplemental table S3.xlsx





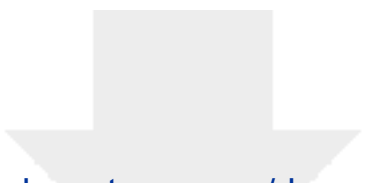
Click here to access/download
Supplementary Material
supplemental table S4.xlsx






Click here to access/download
Supplementary Material
supplementary figure S1.png





Click here to access/download
Supplementary Material
supplemental figure S2.pdf





[Click here to access/download](#)

Supplementary Material

Supplemental file S1. Results of HIFI-barcode.rar

