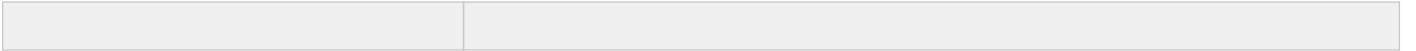


Filling reference gaps via assembling DNA barcodes using high-throughput sequencing - moving toward barcoding the world --Manuscript Draft--

Manuscript Number:	GIGA-D-17-00172R2	
Full Title:	Filling reference gaps via assembling DNA barcodes using high-throughput sequencing - moving toward barcoding the world	
Article Type:	Technical Note	
Funding Information:	the Chinese Universities Scientific Fund (2017QC114)	Not applicable
Abstract:	<p>Over the past decade, biodiversity researchers 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 2000, further dramatic reduction in 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 of 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 added the protocol into protocols.io and integrated the DOI (dx.doi.org/10.17504/protocols.io.ka9csh6) in the revised article. We also have checked and completed the spreadsheet from GigaDB to complete the related GigaDB dataset page.</p>	

	<p>Thank you for all your help and looking forward to hearing from you soon.</p> <p>Sincerely yours, Xin Zhou</p>
Additional Information:	
Question	Response
Are you submitting this manuscript to a special series or article collection?	No
<p>Experimental design and statistics</p> <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>	Yes
<p>Resources</p> <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>	Yes
<p>Availability of data and materials</p> <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 a unique identifier in the references and in the “Availability of Data and Materials” section of your manuscript.</p> <p>Have you have met the above requirement as detailed in our Minimum Standards Reporting Checklist?</p>	Yes



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

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

5
6 1. Beijing Advanced Innovation Center for Food Nutrition and Human Health, College of
7 Plant Protection, China Agricultural University, Beijing 100193, People's Republic of China

8 2. BGI-Shenzhen, Shenzhen, 518083, China

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

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

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

15
16 #Equal contribution

17 *Corresponding to: Xin Zhou, xinzhou@cau.edu.cn

18
19 ORCID details:

20 Shanlin Liu: 0000-0001-8118-8313

21 Chentao Yang: 0000-0003-3447-2316

1 **Abstract**

2

3 Over the past decade, biodiversity researchers have dedicated tremendous
4 efforts in constructing DNA reference barcodes for rapid species registration
5 and identification. Although analytical cost for standard DNA barcoding has
6 been significantly reduced since early 2000, further dramatic reduction in
7 barcoding costs is unlikely because the Sanger sequencing is approaching its
8 limits in throughput and chemistry cost. Constraints in barcoding cost not only
9 led to unbalanced barcoding efforts around the globe, but also refrained
10 High-Throughput-Sequencing (HTS) based taxonomic identification from
11 applying binomial species names, which provide crucial linkages to biological
12 knowledge. We developed an Illumina-based pipeline, HIFI-Barcode, to
13 produce full-length COI barcodes from pooled PCR amplicons generated by
14 individual specimens. The new pipeline generated accurate barcode
15 sequences that were comparable to Sanger standards, even for different
16 haplotypes of the same species that were only a few nucleotides different from
17 each other. Additionally, the new pipeline was much more sensitive in
18 recovering amplicons at low quantity. The HIFI-Barcode pipeline successfully
19 recovered barcodes from over 78% of the PCR reactions that didn't show clear
20 bands on the electrophoresis gel. Moreover, sequencing results based on the
21 single molecular sequencing platform, Pacbio, confirmed the accuracy of the
22 HIFI-Barcode results. Altogether, the new pipeline can provide an improved
23 solution to produce full-length reference barcodes at about 1/10 of the current
24 cost, enabling construction of comprehensive barcode libraries for local fauna,
25 leading to a feasible direction for DNA barcoding global biomes.

26

27

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,
10 e.g., the Barcode of Life Data systems [7], which has recently reached a
11 milestone for 5-million barcodes, covering ca. 0.26 million species (accessed
12 in July 2017). These DNA barcodes have been effectively facilitating species
13 identification, phylogenetic reconstruction [8], and understanding of
14 interspecific interactions 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
18 parks 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 need for biodiversity research are typically suffering from insufficient funding
30 for taxonomy work, especially for DNA based studies. Consequently,

1 HTS-based taxonomic registrations are often constrained to applying
2 Molecular Operational Units (MOTUs) instead of binomial species names,
3 therefore unable to associate existing biological and ecological knowledge to
4 the 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 the costs for sample collection and handling. Further dramatic
10 reductions in barcoding costs is unlikely because Sanger sequencing
11 technology is approaching its limits in throughput and associated chemistry
12 cost. It is estimated that 100 million specimens would need to be sequenced to
13 complete the global barcode registration [1], which translates into a roughly
14 1-billion-dollar budget merely for reference constructions. A similar challenge
15 was seen in the sequencing of the first human genome, where an initial budget
16 of over 3 billion USD was estimated based on the application of Sanger
17 sequencing [19]. Thanks to the advent of HTS technologies over the past
18 decade, the current cost for a human genome is now within the range of a
19 thousand USD, if not 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
23 (e.g., Hiseq and Miseq) have been primarily applied in recent practice [21]; but
24 these are constraint by relatively short read lengths (100-300 bps). Even with
25 the most recent Miseq model at 300bp paired-end (PE) sequencing, full-length
26 barcodes (e.g., ~700 bps for COI including primers) are beyond the
27 sequencing range. Therefore, existing pipelines are forced to produce a
28 fragment of the standard barcodes (e.g., 313bp,[22]) or to apply 2 rounds of
29 PCR amplifications, each targeting on a proportion of the full barcodes [21].
30 Obviously, full-length barcodes are desired for constructing barcode

1 references and extra amplification procedure should be avoided when possible
2 for cost control and simplification of pipelines. In particular, efficient primers
3 might be difficult to identify in the mid-COI barcode region across taxon groups.
4 Alternatively, short HTS reads can be assembled into much longer scaffolds,
5 which is a standard practice in *de novo* genome or transcriptome assembling.
6 In fact, a specific assembly algorithm, SOAPBarcode, has been developed for
7 recovering full-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
21 much more sensitive than Sanger in recovering amplicons at low quantity.
22 Over 78% (25/32) of the “failed” PCR amplicons (those without clear bands on
23 an electrophoresis gel) were successfully recovered at high-quality using the
24 new pipeline. In addition, the single-molecule sequencing platform, Pacbio,
25 has also been adopted in our study to evaluate the accuracy of the
26 HIFI-Barcode method. Altogether, the new pipeline can provide an alternative
27 solution to produce full-length reference barcodes at about 1/10 of the current
28 cost, enabling larger-scale biodiversity barcoding initiatives, especially for
29 areas where DNA 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
6 96-well plates were prepared for the current work: one plate containing 96
7 high-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 of
12 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 ≥ 2 bp differences from each other.
18 Each PCR reaction contained 1 μ l of DNA template, 16.2 μ l of molecular
19 biology grade water, 3 μ l of 10X reaction buffer (Mg^{2+} plus), 2.5 μ l of dNTPs
20 mix (10 mM), 1 μ l of forward and reverse primers (10 mM), and 0.3 μ l of
21 TaKaRa Ex Taq polymerase (5 U/ μ l). The amplification program included a
22 thermocycling profile of 94°C for 1 min, 5 cycles of 94°C for 30 sec, 45°C for 40
23 sec, and an extension at 72°C for 1 min, followed by 35 cycles of 94°C for 30
24 sec, 51°C for 40 sec, and 72°C for 1 min, with a final extension at 72°C for 10
25 min, and finally holding at 12°C. All amplicons were visualized on a 1.2% 96
26 Agarose E-gel (Biowest Agarose). All PCR products from each plate were
27 pooled using 1 μ l per sample resulting in two 96 μ l mixtures, which were sent
28 to BGI and sequenced using a Hiseq 4000. PCR amplicons were fragmented
29 to construct library of an insert-size of 250 bp and sequenced with a strategy of
30 150 PE. A second set of PCR mixture of the 2nd plate (576 μ l, 6 μ l per sample)

1 was also sequenced using PacBio RS II at NextOmics.

3. HIFI-Barcode assembly

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

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

Gap filling: Algorithm adopted from SOAPBarcode [23] (Supplemental Fig. S1) was applied to fill the gaps between 5' and 3' terminal scaffolds of each

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

12 5. Data filtering and read assignment for Pacbio

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

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

24 Barcode sequences obtained by Sanger, HIFI-Barcode method and Pacbio
25 were subject to phylogenetic tree constructions using MEGA7
26 (Neighbor-joining and 1,000 bootstrap) and iTOL [29]. BWA (BWA,
27 RRID:SCR_010910) [30] was applied to align raw reads to assembled HTS
28 barcodes to examine discrepancies between HTS and Sanger sequences.

1 The standard operating procedures are also available from the protocols.io
2 repository [31].

3 **Results**

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

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

15 One cell of Pacbio data containing 28,770 circular consensus sequencing
16 (CCS) from 1,201,158 raw reads were generated for the 2nd plate. CCS reads
17 had an average pass number of 26.5 and were assigned to 82 samples after
18 demultiplexing. (Table 1). Note that a single Pacbio sequencing read can reach
19 as long as 40 kb. Therefore, short CCS read of high quality can be sequenced
20 dozens of times, which in turn effectively corrects sequence errors associated
21 with the platform [32].

22

23 *Accuracy and Efficiency:*

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

1 barcodes were identified to a specific nucleotide in HIFI barcodes (e.g., Fig. 3D,
2 Fig. 4B). Only 9 HIFI barcodes showed a single nucleotide difference from the
3 corresponding Sanger sequences, which could reflect ambiguous base-calling
4 in Sanger sequencing or genuine heteroplasmy in the examined individual. At
5 least two of the discrepancies were proven to be heteroplasmy via mapping
6 raw reads against discrepant sites (Fig. 4A).

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

1 *Non-target sequences detected by HIFI-Barcode*

2 During the HIFI-Barcode assembly procedure, terminus sequences with $\geq 1/10$
3 abundance of that of the most abundant scaffolds at $< 98\%$ similarity were
4 retained for assembly and identity check. This analysis allowed detection of 18
5 non-target sequences co-amplified from the 2 plates (Supplemental Table S4),
6 in addition to COI barcodes. Cross examinations against both NCBI and
7 barcode sequences from the focal plates suggested origins including
8 *Wolbachia* (2), fungus (1), cross-contamination from adjacent wells (7), as well
9 as potential PCR errors and pseudo-genes (8). The presence of non-target
10 PCR products from the 2nd plate were further confirmed by Pacbio sequencing
11 at $> 99\%$ identity, therefore ruling out the likelihood of assembly errors in the
12 HIFI-Barcode pipeline, suggesting they are co-amplified numts present in PCR
13 products. These low-quantity sequences are likely common in regular
14 PCR-based pipelines and detectable by HTS-based approaches. But they can
15 be easily filtered out from genuine COI barcodes following the pipeline
16 described in this study.

17 **Discussion**

18 It is widely acknowledged that we have been undergoing unprecedented
19 global biodiversity loss [33]. DNA-based approaches, e.g., DNA barcoding,
20 DNA metabarcoding, mitochondrial metagenomics (mitochondrial genome
21 skimming), have demonstrated efficacy in accelerating biodiversity inventories
22 of large geographical ranges. These standardized and largely automated
23 procedures will provide pivotal information to understand how biodiversity loss
24 is characterized and how to desist from it. New methodologies enable rapid
25 collection of biodiversity and ecology data at large scale over space and time,
26 which in-turn benefits policy-makers at varied management levels and
27 research groups [34].

28 Interpreting molecular results using existing knowledge on biology, ecology
29 and evolution would require a linkage between DNA references and Linnaeus
30

1 names, which is one of the fundamental roles of DNA barcoding initiatives. The
2 construction of comprehensive barcode references is still, to a large extent,
3 expensive and sometimes prohibitive. This is particularly true for studies
4 targeting on a wide range of taxa from a large area of natural habitat. Although
5 the most represented DNA barcode database (BOLD) now hosts barcodes for
6 0.26 million species, accounting for ca. 1/4 of described species, chances of
7 encountering a novel barcode are still very high, especially for many
8 biodiversity hotspots. Even if an ecological study focuses on just a small
9 proportion of the focal diversity, it is not uncommon that hundreds to thousands
10 of species would need to be barcoded to draw meaningful conclusions. In
11 addition, multiple individuals of the same species (ideally from distinct
12 populations) would need to be sequenced to reflect intraspecific genetic
13 diversities. There is no consensus on the ideal number of conspecific
14 individuals to be sequenced, but in practice an average of 10 is often followed,
15 while some study recommends 20 [35], if not a lot more. Therefore, roughly
16 tens of thousands of individuals, requiring hundreds of thousands of USD, are
17 expected to suffice a regular ecology study, just for the molecular analysis (for
18 a recent example, please see [12]). While the HTS-based approaches have
19 shown promising power in analyzing complex sample mixtures at much
20 reduced unit cost [2-4, 14], one would still need to establish DNA barcode
21 references to be able to go beyond OTU-based interpretation.

22 The HIFI-Barcode method, as the results showed, offers a novel route to
23 produce mass volumes of reliable barcode sequences at significantly reduced
24 cost. The main costs of the HIFI-Barcode pipeline include consumable
25 chemistries, library construction, high-throughput sequencing, and informatics.
26 Despite the increased one-time cost in ordering multiple unique sets of primers,
27 the cost on primers per unit reaction is negligible. Following our protocols, the
28 average cost for a HIFI barcode is around 1 USD, as opposed to 10-20 USD in
29 the standard Sanger approach. Further saving on the production cost is
30 achieved by increased success rates, especially for amplicons with low

1 quantity. In our test, ca. 1/3 of the 2nd plate would have been re-amplified in
2 standard barcoding protocols, using a different set of primers, followed by gel
3 examination, positive picking, PCR purification, and Sanger-sequencing.

4 By complementing the barcode reference library at <1/10 of the current cost,
5 the new approach also reinforces rapid constructions of organelle genomes,
6 e.g., mitochondria and chloroplasts. A number of pilot studies have
7 demonstrated that full mitochondrial genomes can provide elevated power in
8 bulk sample analysis [18, 36]. New approaches to assembling full
9 mito-genomes or the majority of the coding genes have been developed for
10 shotgun sequencing of individual specimens [37], pooled taxa [18, 36], and
11 transcriptomes [37]. In particular, mito-genome assembly through direct
12 shotgun sequencing of mixed taxa can significantly reduce the library
13 construction cost for HTS. Bait sequences, which regularly includes standard
14 COI barcodes, are important for assigning mixed mitochondrial scaffolds to a
15 specific taxon. This is critical especially if the phylogenetic signal of the
16 scaffolds alone is not sufficient to attribute assemblies to species, e.g., when
17 multiple closely related species are pooled. In fact, having multiple bait
18 sequences per species will significantly remove bioinformatics challenge
19 during the assembly procedure [38], which now becomes financially feasible
20 with the help of the HIFI-Barcode pipeline.

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

1 more individuals (e. g. 384-well plate) can further increase the throughput
2 capacity.

3 In summary, the HIFI-Barcode method provides a HTS-based approach with
4 improved economic efficiency, which allows investigators to produce standard
5 full-length barcodes at ca. 1/10 of the current cost. The new protocol not only
6 generates barcode sequences of high quality that are comparable to
7 Sanger-barcodes, but also increases overall sequencing success rates by
8 detecting PCR amplicons in minute quantities. This new method enables
9 construction of comprehensive barcode libraries for local fauna, leading to a
10 feasible direction for DNA barcoding global biomes.

11 **Availability of source code and requirements**

- 12 • Project name: HIFI-Barcode project
- 13 • Project home pages: <https://github.com/comery/HIFI-barcode-hiseq> and
14 <https://github.com/comery/HIFI-barcode-pacbio>
- 15 • Operating system(s): Unix, Linux
- 16 • Programming language: PERL
- 17 • Other requirements: GCC version $\geq 4.4.5$
- 18 • License: GNU General Public License version 3.0 (GPLv3)
- 19 • Any restrictions to use by non-academics: none

20 **Availability of Supporting Data**

21 Supporting snapshots of the HIFI-Barcode code and test
22 data are available in the *GigaScience* GigaDB repository [39]. Raw data and
23 sample information are also available from NCBI bioproject PRJNA414137.
24 The standard operating procedure of HIFI-Barcode is also found in the
25 protocols.io protocols repository [31].

26 **Abbreviations**

27 BOLD - the Barcode of Life Data systems; CCSs – circular consensus

1 sequencing; eDNA – environmental DNA; HTS – High Throughput Sequencing;
2 GB – Gigabase; iDNA – invertebrate-derived DNA; MOTUs – Molecular
3 Operational Units; PCR - Polymerase Chain Reaction; PE – Paired End.
4

5 **Acknowledgements**

6 We thank Dr. Ding Yang from China Agricultural University for contributing
7 samples and Yongfei Du from BGI-Shenzhen for his help in coordinating the
8 HiSeq sequencing of our 2nd plate. We would like to thank Min Tang, Guanliang
9 Meng for assistant in figure preparation. This work is supported by the China
10 National GeneBank, BGI and by China Agricultural University through the
11 Chinese Universities Scientific Fund, 2017QC114 to XZ.
12

13 **Author contributions:**

14 X.Z. and S.L. designed the study; S.L. coordinated the project and led the
15 analyses; C.Z. and C.Y. led the bench work and contributed to the analyses;
16 S.L., C.Z., and C.Y. formulated the early drafts, and X.Z. revised the
17 manuscript.
18

19 **Competing interests**

20 The authors declare that they have no competing interests.
21

22 **References**

- 23
- 24 1. Hebert PD, Hollingsworth PM, Hajibabaei M: **From writing to reading the**
25 **encyclopedia of life**. *Phil Trans R Soc B* 2016, **371**(1702):20150321.
- 26 2. Yu DW, Ji YQ, Emerson BC, Wang XY, Ye CX, Yang CY, Ding ZL: **Biodiversity Soup:**
27 **Metabarcoding of arthropods for rapid biodiversity assessment and**
28 **biomonitoring**. *Methods in Ecology and Evolution* 2012, **3**(4):613-623.
- 29 3. Zhou X, Li Y, Liu S, Yang Q, Su X, Zhou L, Tang M, Fu R, Li J, Huang Q: **Ultra-deep**
30 **sequencing enables high-fidelity recovery of biodiversity for bulk arthropod**
31 **samples without PCR amplification**. *GigaScience* 2013, **2**(1):4.
- 32 4. Taberlet P, Coissac E, Hajibabaei M, Rieseberg LH: **Environmental DNA**. *Molecular*
33 *Ecology* 2012, **21**:1789–1793.

- 1 5. Calvignac - Spencer S, Merkel K, Kutzner N, Kühl H, Boesch C, Kappeler PM,
2 Metzger S, Schubert G, Leendertz FH: **Carrion fly - derived DNA as a tool for**
3 **comprehensive and cost - effective assessment of mammalian biodiversity.**
4 *Molecular Ecology* 2013, 22(4):915-924.
- 5 6. Schnell IB, Thomsen PF, Wilkinson N, Rasmussen M, Jensen LR, Willerslev E,
6 Bertelsen MF, Gilbert MTP: **Screening mammal biodiversity using DNA from**
7 **leeches.** *Current biology* 2012, 22(8):R262-R263.
- 8 7. Ratnasingham S, Hebert PDN: **BOLD: The Barcode of Life Data System**
9 **(<http://www.barcodinglife.org>).** *Molecular ecology notes* 2007, 7(3):355-364.
- 10 8. Zhou X, Frandsen PB, Holzenthal RW, Beet CR, Bennett KR, Blahnik RJ, Bonada N,
11 Cartwright D, Chuluunbat S, Cocks GV: **The Trichoptera barcode initiative: a**
12 **strategy for generating a species-level Tree of Life.** *Phil Trans R Soc B* 2016,
13 **371(1702):20160025.**
- 14 9. Miller SE, Hausmann A, Hallwachs W, Janzen DH: **Advancing taxonomy and**
15 **bioinventories with DNA barcodes.** *Phil Trans R Soc B* 2016, **371(1702):20150339.**
- 16 10. Zhou X, Robinson JL, Geraci CJ, Parker CR, Flint Jr OS, Etnier DA, Ruiter D, DeWalt
17 RE, Jacobus LM, Hebert PD: **Accelerated construction of a regional DNA-barcode**
18 **reference library: caddisflies (Trichoptera) in the Great Smoky Mountains**
19 **National Park.** *Journal of the North American Benthological Society* 2011,
20 **30(1):131-162.**
- 21 11. Zhou X, Adamowicz SJ, Jacobus LM, DeWalt RE, Hebert PD: **Towards a**
22 **comprehensive barcode library for arctic life-Ephemeroptera, Plecoptera, and**
23 **Trichoptera of Churchill, Manitoba, Canada.** *Frontiers in zoology* 2009, 6(1):1.
- 24 12. Telfer AC, Young MR, Quinn J, Perez K, Sobel CN, Sones JE, Levesque-Beaudin V,
25 Derbyshire R, Fernandez-Triana J, Rougerie R: **Biodiversity inventories in high**
26 **gear: DNA barcoding facilitates a rapid biotic survey of a temperate nature**
27 **reserve.** *Biodiversity data journal* 2015(3).
- 28 13. Valentini A, Taberlet P, Miaud C, Civade R, Herder J, Thomsen PF, Bellemain E,
29 Besnard A, Coissac E, Boyer F: **Next - generation monitoring of aquatic**
30 **biodiversity using environmental DNA metabarcoding.** *Molecular Ecology* 2016.
- 31 14. Hajibabaei M, Shokralla S, Zhou X, Singer GAC, Baird DJ: **Environmental barcoding:**
32 **a next-generation sequencing approach for biomonitoring applications using**
33 **river benthos.** *PloS one* 2011, 6(4):e17497.
- 34 15. Coissac E, Riaz T, Puillandre N: **Bioinformatic challenges for DNA metabarcoding**
35 **of plants and animals.** *Molecular Ecology* 2012, 21:1834–1847.
- 36 16. Bienert F, Dedanieli S, Miquel C, Coissac E, Poillot C, Brun JJ, Taberlet P: **Tracking**
37 **earthworm communities from soil DNA.** *Molecular Ecology* 2012, 21(8):2017-2030.
- 38 17. Baird DJ, Hajibabaei M: **Biomonitoring 2.0: a new paradigm in ecosystem**
39 **assessment made possible by next-generation DNA sequencing..** *Molecular*
40 *Ecology* 2012, 21(8):2039-2044.
- 41 18. Gillett CP, Crampton-Platt A, Timmermans MJ, Jordal B, Emerson BC, Vogler AP:
42 **Bulk de novo mitogenome assembly from pooled total DNA elucidates the**
43 **phylogeny of weevils (Coleoptera: Curculionoidea).** *Molecular biology and*
44 *evolution* 2014:msu154.

- 1 19. Krulwich R, LANDER E: **Cracking the Code of Life**. In.: *Public Broadcasting Service*;
2 2001.
- 3 20. Shokralla S, Gibson JF, Nikbakht H, Janzen DH, Hallwachs W, Hajibabaei M: Next -
4 4 generation DNA barcoding: **using next - generation sequencing to enhance and**
5 5 **accelerate DNA barcode capture from single specimens**. *Molecular Ecology*
6 6 *Resources* 2014, 14(5):892-901.
- 7 21. Shokralla S, Porter TM, Gibson JF, Dobosz R, Janzen DH, Hallwachs W, Golding GB,
8 8 Hajibabaei M: **Massively parallel multiplex DNA sequencing for specimen**
9 9 **identification using an Illumina MiSeq platform**. *Scientific reports* 2015, 5.
- 10 22. Meier R, Wong W, Srivathsan A, Foo M: **\$1 DNA barcodes for reconstructing**
11 11 **complex phenomes and finding rare species in specimen - rich samples**.
12 12 *Cladistics* 2016, 32(1):100-110.
- 13 23. Liu S, Li Y, Lu J, Su X, Tang M, Zhang R, Zhou L, Zhou C, Yang Q, Ji Y:
14 14 **SOAPBarcode: revealing arthropod biodiversity through assembly of Illumina**
15 15 **shotgun sequences of PCR amplicons**. *Methods in Ecology and Evolution* 2013,
16 16 4(12):1142-1150.
- 17 24. Ivanova NV, Dewaard JR, Hebert PD: **An inexpensive, automation - friendly**
18 18 **protocol for recovering high - quality DNA**. *Molecular Ecology Resources* 2006,
19 19 6(4):998-1002.
- 20 25. Vrijenhoek R: **DNA primers for amplification of mitochondrial cytochrome c**
21 21 **oxidase subunit I from diverse metazoan invertebrates**. *Molecular marine biology*
22 22 *and biotechnology* 1994, 3(5):294-299.
- 23 26. Rognes T, Flouri T, Nichols B, Quince C, Mahé F: **VSEARCH: a versatile open**
24 24 **source tool for metagenomics**. *PeerJ* 2016, 4:e2584.
- 25 27. Liu B, Yuan J, Yiu SM, Li Z, Xie Y, Chen Y, Shi Y, Zhang H, Li Y, Lam TW: **COPE: an**
26 26 **accurate k-mer-based pair-end reads connection tool to facilitate genome**
27 27 **assembly**. *Bioinformatics* 2012, 28(22):2870-2874.
- 28 28. Pacbio SmrtAnalysis pipeline <https://github.com/PacificBiosciences>
- 29 29. Letunic I, Bork P: **Interactive tree of life (iTOL) v3: an online tool for the display**
30 30 **and annotation of phylogenetic and other trees**. *Nucleic Acids Research*
31 31 2016:gkw290.
- 32 30. Li H, Durbin R: **Fast and accurate short read alignment with Burrows–Wheeler**
33 33 **transform**. *Bioinformatics* 2009, 25(14):1754-1760.
- 34 31. Shanlin Liu, Chentao Yang, Chengran Zhou, Xin Zhou (1508072650) (2017). Filling
35 35 reference gaps via assembling DNA barcodes using high-throughput sequencing –
36 36 moving toward barcoding the world — HIFI-Barcode SOP. **protocols.io**
37 37 <http://dx.doi.org/10.17504/protocols.io.ka9csh6>
- 38 32. Eid J, Fehr A, Gray J, Luong K, Lyle J, Otto G, Peluso P, Rank D, Baybayan P,
39 39 Bettman B: **Real-time DNA sequencing from single polymerase molecules**.
40 40 *Science* 2009, 323(5910):133-138.
- 41 33. Pimm SL, Jenkins CN, Abell R, Brooks TM, Gittleman JL, Joppa LN, Raven PH,
42 42 Roberts CM, Sexton JO: **The biodiversity of species and their rates of extinction,**
43 43 **distribution, and protection**. *Science* 2014, 344(6187):1246752.
- 44 34. Díaz S, Demissew S, Carabias J, Joly C, Lonsdale M, Ash N, Larigauderie A, Adhikari

1 JR, Arico S, Báldi A: **The IPBES Conceptual Framework—connecting nature and**
2 **people.** *Current Opinion in Environmental Sustainability* 2015, **14**:1-16.

3 35. Luo A, Lan H, Ling C, Zhang A, Shi L, Ho SY, Zhu C: **A simulation study of sample**
4 **size for DNA barcoding.** *Ecology and Evolution* 2015, **5**(24):5869-5879.

5 36. Tang M, Tan M, Meng G, Yang S, Su X, Liu S, Song W, Li Y, Wu Q, Zhang A:
6 **Multiplex sequencing of pooled mitochondrial genomes—a crucial step toward**
7 **biodiversity analysis using mito-metagenomics.** *Nucleic Acids Research* 2014,
8 **42**(22):e166-e166.

9 37. Tan M, Zhang R, Xiang C, Zhou X: **The complete mitochondrial genome of the**
10 **pumpkin fruit fly, *Bactrocera tau* (Diptera: Tephritidae).** *Mitochondrial DNA Part A*
11 2016, **27**(4):2502-2503.

12 38. Timmermans MJ, Dodsworth S, Culverwell C, Bocak L, Ahrens D, Littlewood DT, Pons
13 J, Vogler AP: **Why barcode? High-throughput multiplex sequencing of**
14 **mitochondrial genomes for molecular systematics.** *Nucleic acids research* 2010,
15 **38**(21):e197-e197.

16 39. Liu, S; Yang, C; Zhou, C; Zhou, X (2017): **Supporting data for "Filling reference**
17 **gaps via assembling DNA barcodes using high-throughput sequencing -**
18 **moving toward barcoding the world"** GigaScience Database.
19 <http://dx.doi.org/10.5524/100363>

20

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 Figure S3. PCR electrophoresis results of the 2nd plate.

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

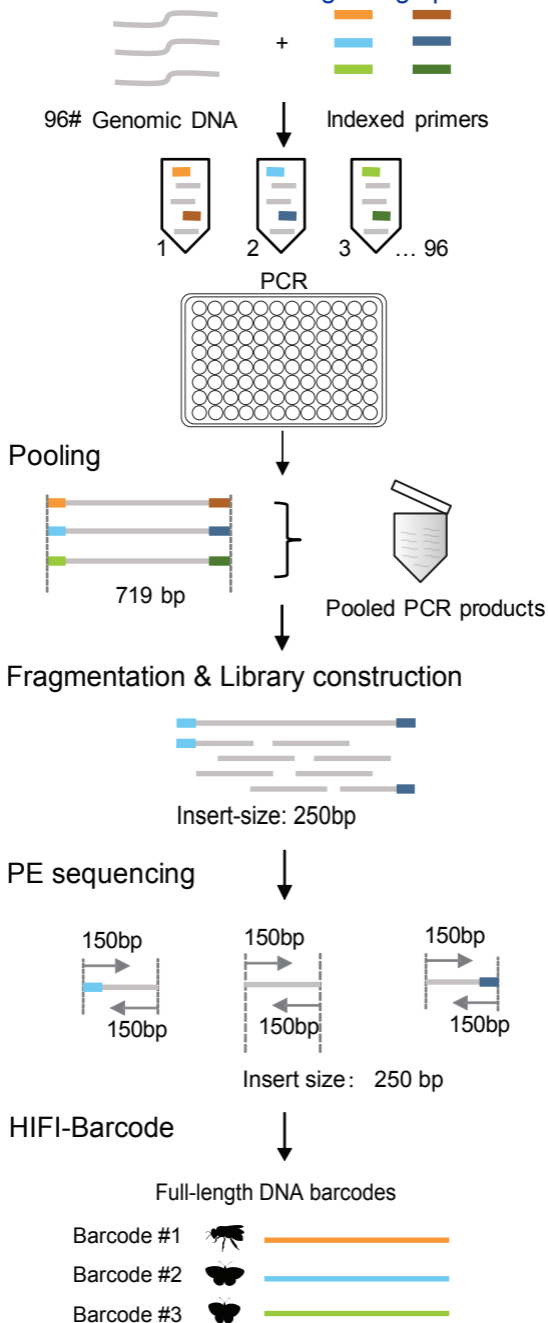
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

Supplemental File S1. Results of HIFI-barcode

Figure 1

Indexed PCR

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



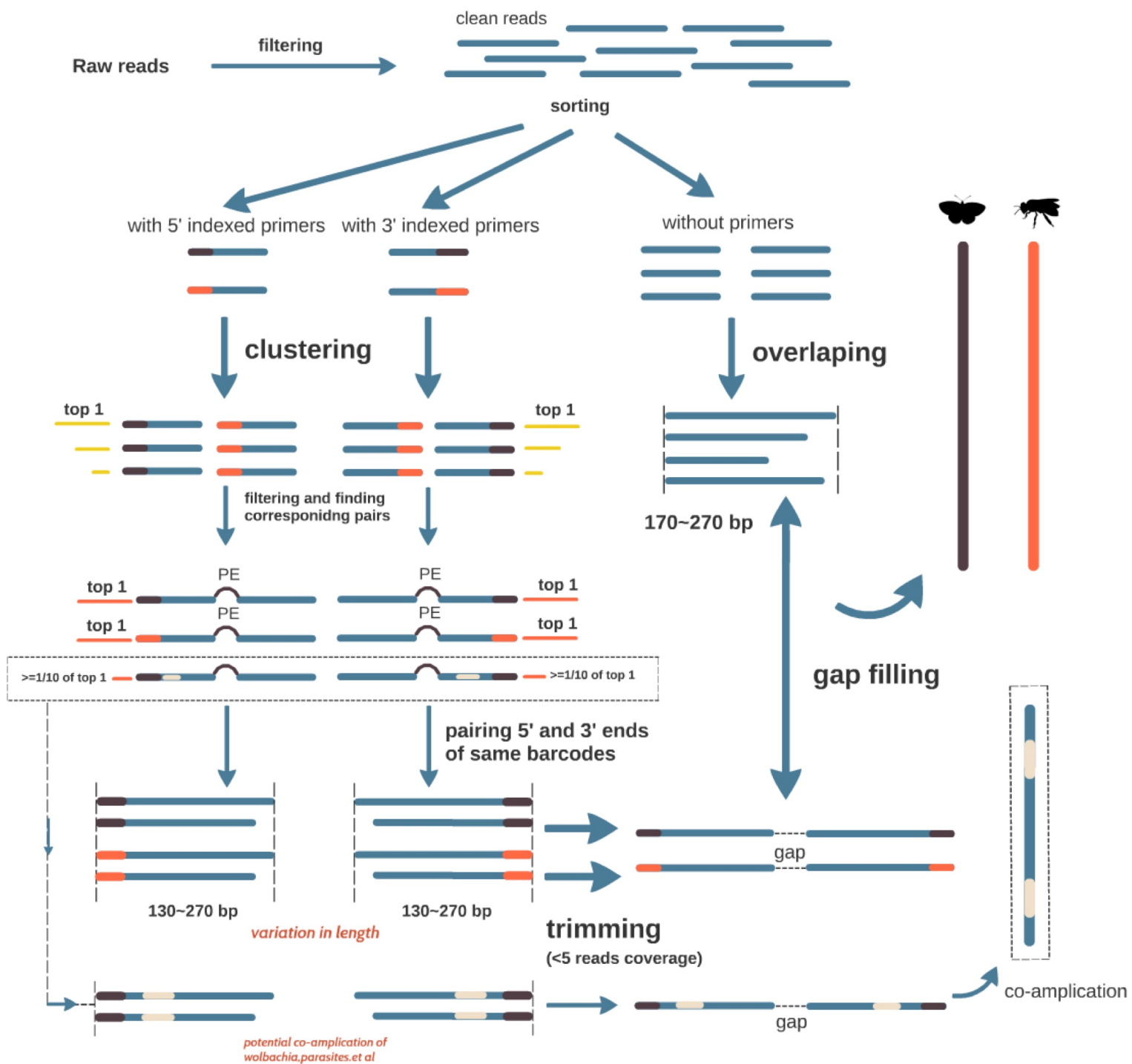
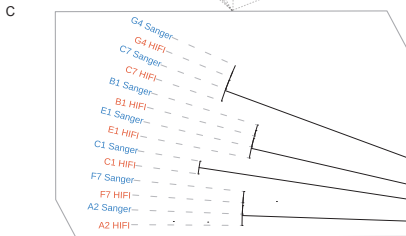
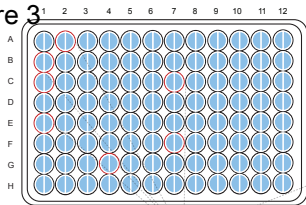


Figure 3



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



Tree scale
0.01

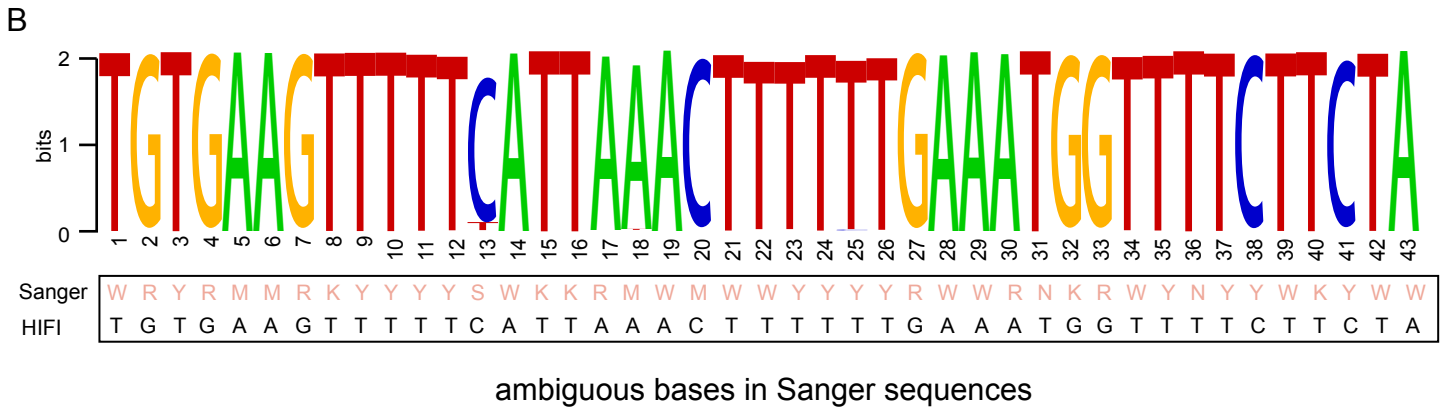
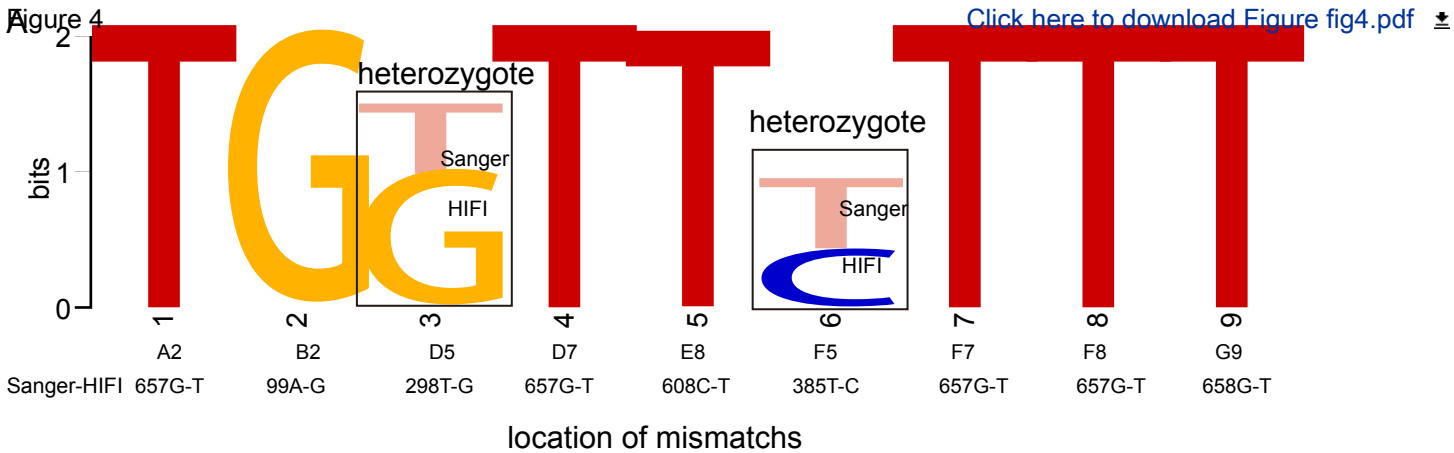
D

Neptis sappho:

B6_Sanger **t**atw**tt**att**aa**cagat**cg**gaat**mw**waata**cc**yy**ca**ttttt**tg**ay**cc**tg**ct**gragg**ggg**
 B6_HIFI **t**ata**tt**att**aa**cagat**cg**gaat**ct**taata**ct**tt**ca**ttttt**tg**at**cc**tg**ct**ggagg**ggg**

Aporia genestieri:

B1_Sanger **a**ca**r**ta**t**acc**cc**ctc**t**tt**ct**ct**ta**at**at**kg**ct**ca**gg**agg**ct**ct**ct**gt**ag**at**tt**a
 B1_HIFI **a**ca**g**ta**t**acc**cc**ctc**t**tt**ct**ct**ta**at**at**tg**ct**ca**gg**agg**ct**ct**ct**gt**ag**at**tt**a





Click here to access/download
Supplementary Material
supplementary table S1.xlsx





Click here to access/download
Supplementary Material
supplementary table S2.xlsx



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





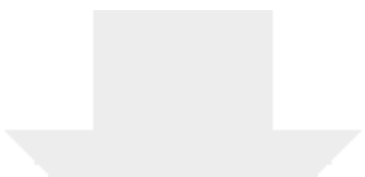
Click here to access/download
Supplementary Material
supplemental table S4.xlsx






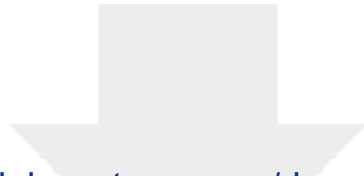
Click here to access/download
Supplementary Material
supplementary figure S1.png





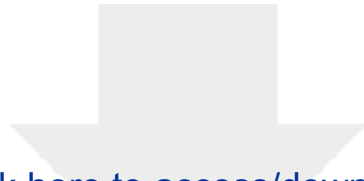
Click here to access/download
Supplementary Material
supplemental figure S2.pdf





Click here to access/download
Supplementary Material
Supplemental Figure S3.tif





[Click here to access/download](#)

Supplementary Material

Supplemental file S1. Results of HIFI-barcode.rar

