

GigaScience

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

Manuscript Number:	GIGA-D-17-00172	
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:		
Opposed Reviewers:	<p>Mehrdad Hajibabae, Ph.D. Biodiversity Institute of Ontario & Integrative Biology, University of Guelph mhajibab@uoguelph.ca Direct competitor on barcode method development. We request you exclude anyone from this institute from reviewing the manuscript</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

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

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

1. Beijing Advanced Innovation Center for Food Nutrition and Human Health, China Agricultural University, Beijing 100193, People's Republic of China
2. BGI-Shenzhen, Shenzhen, 518083, China
3. College of Plant Protection, China Agricultural University, Beijing 100193, People's Republic of China
4. Centre for GeoGenetics, Natural History Museum of Denmark, University of Copenhagen, Copenhagen, Denmark
5. Key Laboratory of Bio-Resources and Eco-Environment, Ministry of Education, College of Life Sciences, Sichuan University, Chengdu 610000, China.

#Equal contribution

*Corresponding to: Xin Zhou

1
2
3
4
5
6
7
8
9
10
11
12
13
14
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

Abstract

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.

Background

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

Along with the rapid accumulation of global barcode references for various taxon groups, significant effort has been made in digitalizing biomes, e.g., sequencing all taxa of particular lineages found in entire range of national parks or islands [9]. Early efforts in barcoding biomes have employed standard Sanger sequencing-based approaches to characterizing focal fauna [10-12]. Alternatively, boosted by HTS technologies, DNA metabarcoding and mitochondrial metagenomics (mitochondrial genome skimming) have been applied in investigations of local biodiversity and in evaluation of biological managements [13-17]. These practices allow investigators to rapidly understand species richness or even approximation for species evenness and/or biomass for complex biological samples [4, 18]. A typical dilemma, however, is the lack of local barcode references, from which HTS biodiversity analysis could draw conclusions on species occurrences. This is primarily due to unbalanced barcoding efforts around the globe, where regions in desperate needs for biodiversity research are typically suffering from insufficient funding 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
6

7
8 Admittedly, the analytical cost for standard DNA barcoding has been
9 significantly reduced since early 2000, a result from the development of
10 centralized and industrialized barcoding facilities and automated pipelines [1].
11 Currently, the average production cost for a reference barcode is ca. 10 USD,
12 excluding that for sample collection and handling. Further dramatic reduction
13 on barcoding costs is unlikely because Sanger sequencing technology is
14 approaching its limits in throughput and associated chemistry cost. It is
15 estimated that 100 million specimens would need to be sequenced to complete
16 the global barcode registration [1], which translates into a roughly 1-billion-
17 dollar budget for merely reference constructions. A similar challenge was seen
18 in the sequencing of the first human genome, where an initial budget of over 3
19 billion USD was estimated based on the application of Sanger sequencing [19].
20 Thanks to the advent of HTS technologies over the past decade, the current
21 cost for a human genome is within the range of just a few thousand USD, if not
22 less.
23
24

25
26 An early study using HTS in generating barcodes from single specimens
27 employed the Roche 454 platform [20], which was rapidly phased out due to
28 limited throughput capacity (hence high chemistry cost). Illumina platforms (e.g.,
29 Hiseq and Miseq) have been primarily applied in recent practices [21]; but these
30 are constraint by relatively short read lengths (100-300 bps). Even with the most
31 recent Miseq model at 300bp paired-end (PE) sequencing, full-length barcodes
32 (e.g., ~700 bps for COI including primers) are beyond the sequencing range.
33 Therefore, existing pipelines are forced to produce a fragment of the standard
34 barcodes (e.g., 313bp,[22]) or to apply 2 rounds of PCR amplifications, each
35 targeting on a proportion of the full barcodes [21]. Obviously, full-length
36 barcodes are desired for constructing barcode references and extra
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 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 in recovering amplicons at low quantity. Over 78% (25/32) of the
22 “failed” PCR amplicons (those without clear bands on an electrophoresis gel)
23 were successfully recovered at high-quality using the new pipeline. In addition,
24 the single-molecule sequencing platform, Pacbio, has also been adopted in our
25 study to evaluate the accuracy of the HIFI-Barcode method. Altogether, the new
26 pipeline can provide an alternative solution to produce full-length reference
27 barcodes at about 1/10 of the current cost, enabling larger-scale biodiversity
28 barcoding initiatives, especially for areas where DNA references are scarce.

29 **Material and methods**

1. DNA preparation

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

2. DNA amplification and sequencing

Ninety-six pairs of different tags were added to both ends of a common COI barcode primer set (LCO1490 and HCO2198, Supplemental Table S1) [25], with each tag containing 5 bps allowing for ≥ 3 bp differences from each other. Each PCR reaction contained 1 μ l of DNA template, 16.2 μ l of molecular biology grade water, 3 μ l of 10X reaction buffer (Mg²⁺ plus), 2.5 μ l of dNTPs mix (10 mM), 1 μ l of forward and reverse primers (10 mM), and 0.3 μ l of TaKaRa Ex Taq polymerase (5 U/ μ l). The amplification program included a touchdown thermocycling profile of 94°C for 1 min, 5 cycles of 94°C for 30 sec, 45°C for 40 sec, and an extension at 72°C for 1 min, followed by 35 cycles of 94°C for 30 sec, 51°C for 40 sec, and 72°C for 1 min, with a final extension at 72°C for 10 min, and finally holding at 12°C. All amplicons were visualized on a 1.2% 96 Agarose E-gel (Biowest Agarose). All PCR products from each plate were pooled using 1 μ l per sample resulting in two 96 μ l mixtures, which were sent to BGI and sequenced using a Hiseq 4000. PCR amplicons were fragmented to construct library of an insert-size of 250 bp and sequenced with a strategy of 150 PE. A second set of PCR mixture of the 2nd plate (576 μ l, 6 μ l per sample) was 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. 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, 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 [26] 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 individual to complete the full-length barcodes. Briefly, for each individual, the 5' end was defined as the start point, and the 3' end as the end point. Then, the kmer set from de brujin graph was walked step by step from the start point to

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

18 5. Data filtering and read assignment for Pacbio

20 Pacbio SmrtAnalysis pipeline (<https://github.com/PacificBiosciences>) was
21 adopted to extract 28,770 circular consensus sequencing (CCSs) from 1.1G
22 raw data. Then, CCSs of ≥ 15 passes were chosen for next steps: 1) 22,075
23 CCSs were demultiplexed by their corresponding indices using an in-house Perl
24 script, allowing a maximum of 1bp deletion at the 5' end of forward index or the
25 3' end of reverse index. 2) for each sample, sequences with a length range out
26 of 658 ± 6 bp were removed and the remaining unique sequences were sorted
27 by pass numbers and identical sequences were clustered together; 3) unique
28 sequence of the most abundant cluster was retained as the correct barcode
29 sequence for each sample.
30
31
32
33
34
35
36
37
38
39
40
41
42
43

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

45 Barcode sequences obtained by Sanger, HIFI-Barcode method and Pacbio
46 were subject to phylogenetic tree constructions using MEGA7 (Neighbor-joining
47 and 1,000 bootstrap) and iTOL [28]. BWA [29] was applied to align raw reads
48 to assembled HTS barcodes to examine discrepancies between HTS and
49 Sanger sequences.
50
51
52
53
54
55
56
57

58 **Results**

59 A total of 4,824,443 and 4,439,345 PE reads for the 1st and 2nd plate were
60
61
62
63
64
65

1 obtained after data filtering, respectively, using Hiseq 4000.

2 For the 1st plate, a total of 1,910,616 (39.60%) reads were assigned to their
3 corresponding samples as either 5' or 3' end, and 1,898,372 (39.34%) as reads
4 belonging to intermediate regions, while 1,015,455 (21.05%) reads were
5 identified as primer dimers or short PCR chimeras. The abundance of end
6 reads for each sample varies significantly, ranging from 2,444 to 64,705. After
7 clustering at 100%-similarity for the 5' and 3' end reads, most samples (61 out
8 of 96) obtained single unique reads after read assignment. The 2nd plate
9 possessed similar read distribution and details of both plates were summarized
10 in Table 1.
11
12
13
14
15
16
17
18
19

20
21 One cell of Pacbio data containing 28,770 circular consensus sequencing (CCS)
22 from 1,201,158 raw reads were generated for the 2nd plate. CCS reads had an
23 average pass number of 26.5 and were assigned to 82 samples after
24 demultiplexing. (Table 1). Note that a single Pacbio sequencing read can reach
25 as long as 40 kb. Therefore, short CCS read of high quality can be sequenced
26 dozens of times, which in turn effectively corrects sequence errors associated
27 with the platform [30].
28
29
30
31
32
33
34
35
36

37 *Accuracy and Efficiency:*

38
39 Sanger barcodes were obtained from all 96 lepidopteran samples of the 1st
40 plate (Fig. 3A), including 91 haplotypes and 85 OTUs using a similarity
41 threshold of $\geq 98\%$. The HIFI-Barcode assemblies were successful for all 96
42 samples and showed high accuracy compared to Sanger sequences. Even
43 identical or highly similar barcodes from individuals of the same species were
44 correctly assembled, e.g., A2 versus F7, B1 versus E1, and C7 versus G4 (Fig.
45 3B and 3C). A total of 43 ambiguous sites (out of 63,168 bps) found in Sanger
46 barcodes were identified to a specific nucleotide in HIFI barcodes (e.g., Fig. 3D,
47 Fig. 4B). Only 9 HIFI barcodes showed a single nucleotide difference from the
48 corresponding Sanger sequences, which could reflect ambiguous base-calling
49 in Sanger sequencing or genuine heteroplasmy in the examined individual. At
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

1 least two of the discrepancies were proven to be heterozygotes via mapping
2 raw reads against discrepant sites (Fig. 4A).
3

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

52 During the HIFI-Barcode assembly procedure, terminus sequences with $\geq 1/10$
53 abundance of that of the most abundant scaffolds at <98% similarity were
54 retained for assembly and identity check. This analysis allowed detection of 18
55 non-target sequences co-amplified from the 2 plates (Supplemental Table S4),
56
57
58
59
60
61
62
63
64
65

1 in addition to COI barcodes. Cross examinations against both NCBI and
2 barcode sequences from the focal plates suggested origins including
3 *Wolbachia* (2), fungus (1), cross-contamination from adjacent wells (7), as well
4 as potential PCR errors and pseudo-genes (8). The presence of non-target
5 PCR products from the 2nd plate were further confirmed by Pacbio sequencing
6 at >99% identity, therefore ruling out the likelihood of assembly errors in the
7 HIFI-Barcode pipeline. These low-quantity sequences are likely common in
8 regular PCR-based pipelines and detectable by HTS-based approaches. But
9 they can be easily filtered out from genuine COI barcodes following the pipeline
10 described in this study.
11
12
13
14
15
16
17
18
19
20
21
22

23 **Discussion**

24 It is widely acknowledged that we have been undergoing unprecedented global
25 biodiversity loss [31]. DNA-based approaches, e.g., DNA barcoding, DNA
26 metabarcoding, mitochondrial metagenomics (mitochondrial genome
27 skimming), have demonstrated efficacy in accelerating biodiversity inventories
28 of large geographical ranges. These standardized and largely automated
29 procedures will provide pivotal information to understand how biodiversity loss
30 is characterized and how to desist from it. New methodologies enable rapid
31 collection of biodiversity and ecology data at large scale over space and time,
32 which in-turn benefits policy-makers at varied management levels and research
33 groups [32].
34
35
36
37
38
39
40
41
42
43
44

45 Interpreting molecular results using existing knowledge on biology, ecology and
46 evolution would require a linkage between DNA references and Linnaeus
47 names, which is one of the fundamental roles of DNA barcoding initiatives. The
48 construction of comprehensive barcode references is still, to a large extent,
49 expensive and sometimes prohibitive. This is particularly true for studies
50 targeting on a wide range of taxa from a large area of natural habitat. Although
51 the most represented DNA barcode database (BOLD) now hosts barcodes for
52 0.26 million species, accounting for ca. 1/4 of described species, chances of
53
54
55
56
57
58
59
60
61
62
63
64
65

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

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

28 By complementing the barcode reference library at <1/10 of the current cost,
29 the new approach also reinforces rapid constructions of organelle genomes,
30 e.g., mitochondria and chloroplasts. A number of pilot studies have
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 demonstrated that full mitochondrial genomes can provide elevated power in
2 bulk sample analysis [18, 34]. New approaches to assembling full mito-
3 genomes or the majority of the coding genes have been developed for shotgun
4 sequencing of individual specimens [35], pooled taxa [18, 34], and
5 transcriptomes [35]. In particular, mito-genome assembly through direct
6 shotgun sequencing of mixed taxa can significantly reduce the library
7 construction cost for HTS. Bait sequences, which regularly includes standard
8 COI barcodes, are important for assigning mixed mitochondrial scaffolds to a
9 specific taxon. This is critical especially if the phylogenetic signal of the
10 scaffolds alone is not sufficient to attribute assemblies to species, e.g., when
11 multiple closely related species are pooled. In fact, having multiple bait
12 sequences per species will significantly remove bioinformatics challenge during
13 the assembly procedure [36], which now becomes financially feasible with the
14 help of the HIFI-Barcode pipeline.
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30

31 Several aspects can be further improved for our method: 1) Multiple barcode
32 markers (e.g., COI, CYTB, 12S, etc.) can be pooled in a single shotgun
33 sequencing effort without increasing tag complexity, which will again alleviate
34 analytical cost. 2) The pooled PCR amplicons were subject to library
35 construction directly in the present study. The proportion of primer dimers and
36 short PCR chimera reached as high as ca. 21% in our raw reads, which can be
37 easily removed using size-preference magnetic beads. 3) Addition of inosine to
38 the 3' terminus of the primer may increase its universality and will further elevate
39 the successful rate and efficiency. 4) Longer tags allowing for pooling more
40 individuals (e. g. 384-well plate) can further increase the throughput capacity.
41
42
43
44
45
46
47
48
49
50

51 In summary, the HIFI-Barcode method provides a HTS-based approach with
52 improved economic efficiency, which allows investigators to produce standard
53 full-length barcodes at ca. 1/10 of the current cost. The new protocol not only
54 generates barcode sequences of high quality that are comparable to Sanger-
55 barcodes, but also increases overall sequencing success rates by detecting
56
57
58
59
60
61
62
63
64
65

1 PCR amplicons in minute quantities. This new method enables construction of
2 comprehensive barcode libraries for local fauna, leading to a feasible direction
3 for DNA barcoding global biomes.
4
5
6
7

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

15 **Availability of Supporting Data**

16 Availability of supporting data and materials code and data are available in the
17 GigaScience GigaDB repository (XXX). Source code also can be found in
18 <https://github.com/comery/HIFI-barcode-hiseq> and
19 <https://github.com/comery/HIFI-barcode-pacbio>.
20
21
22
23
24

25 **Abbreviations**

26 BOLD - the Barcode of Life Data systems; CCSs – circular consensus
27 sequencing; eDNA – environmental DNA; HTS – High Throughput Sequencing;
28 GB – Gigabase; iDNA – invertebrate-derived DNA; MOTUs – Molecular
29 Operational Units; PCR - Polymerase Chain Reaction; PE – Paired End.
30
31
32
33
34
35
36

37 **Acknowledgements**

38 We thank Dr. Ding Yang from China Agricultural University for contributing
39 samples and Yongfei Du from BGI-Shenzhen for his help on coordinating the
40 HiSeq sequencing of our 2nd plate. We would like to thank Min Tang, Guanliang
41 Meng for assistant in figure preparation. This work is supported by the China
42 National GeneBank, BGI and by China Agricultural University through the
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

Chinese Universities Scientific Fund, 2017QC114 to XZ.

Author contributions:

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

Competing interests

The authors declare that they have no competing interests.

1. Hebert PD, Hollingsworth PM, Hajibabaei M: **From writing to reading the encyclopedia of life**. *Phil Trans R Soc B* 2016, **371**(1702):20150321.
2. Yu DW, Ji YQ, Emerson BC, Wang XY, Ye CX, Yang CY, Ding ZL: **Biodiversity Soup: Metabarcoding of arthropods for rapid biodiversity assessment and biomonitoring**. *Methods in Ecology and Evolution* 2012, **3**(4):613-623.
3. Zhou X, Li Y, Liu S, Yang Q, Su X, Zhou L, Tang M, Fu R, Li J, Huang Q: **Ultra-deep sequencing enables high-fidelity recovery of biodiversity for bulk arthropod samples without PCR amplification**. *GigaScience* 2013, **2**(1):4.
4. Taberlet P, Coissac E, Hajibabaei M, Rieseberg LH: **Environmental DNA**. *Molecular Ecology* 2012, **21**:1789–1793.
5. Calvignac-Spencer S, Merkel K, Kutzner N, Kühl H, Boesch C, Kappeler PM, Metzger S, Schubert G, Leendertz FH: **Carrion fly-derived DNA as a tool for comprehensive and cost-effective assessment of mammalian biodiversity**. *Molecular Ecology* 2013, **22**(4):915-924.
6. Schnell IB, Thomsen PF, Wilkinson N, Rasmussen M, Jensen LR, Willerslev E, Bertelsen MF, Gilbert MTP: **Screening mammal biodiversity using DNA from leeches**. *Current biology* 2012, **22**(8):R262-R263.
7. Ratnasingham S, Hebert PDN: **BOLD: The Barcode of Life Data System (<http://www.barcodinglife.org>)**. *Molecular ecology notes* 2007, **7**(3):355-364.
8. Zhou X, Frandsen PB, Holzenthal RW, Beet CR, Bennett KR, Blahnik RJ, Bonada N, Cartwright D, Chuluunbat S, Cocks GV: **The Trichoptera barcode initiative: a strategy for generating a species-level Tree of Life**. *Phil Trans R Soc B* 2016, **371**(1702):20160025.
9. Miller SE, Hausmann A, Hallwachs W, Janzen DH: **Advancing taxonomy and bioinventories with DNA barcodes**. *Phil Trans R Soc B* 2016, **371**(1702):20150339.
10. Zhou X, Robinson JL, Geraci CJ, Parker CR, Flint Jr OS, Etnier DA, Ruitter D, DeWalt

1 RE, Jacobus LM, Hebert PD: **Accelerated construction of a regional DNA-barcode**
2 **reference library: caddisflies (Trichoptera) in the Great Smoky Mountains**
3 **National Park.** *Journal of the North American Benthological Society* 2011, **30**(1):131-
4 162.

- 5 11. Zhou X, Adamowicz SJ, Jacobus LM, DeWalt RE, Hebert PD: **Towards a**
6 **comprehensive barcode library for arctic life-Ephemeroptera, Plecoptera, and**
7 **Trichoptera of Churchill, Manitoba, Canada.** *Frontiers in zoology* 2009, **6**(1):1.
8
9 12. Telfer AC, Young MR, Quinn J, Perez K, Sobel CN, Sones JE, Levesque-Beaudin V,
10 Derbyshire R, Fernandez-Triana J, Rougerie R: **Biodiversity inventories in high gear:**
11 **DNA barcoding facilitates a rapid biotic survey of a temperate nature reserve.**
12 *Biodiversity data journal* 2015(3).
13
14 13. Valentini A, Taberlet P, Miaud C, Civade R, Herder J, Thomsen PF, Bellemain E,
15 Besnard A, Coissac E, Boyer F: **Next-generation monitoring of aquatic biodiversity**
16 **using environmental DNA metabarcoding.** *Molecular Ecology* 2016.
17
18 14. Hajibabaei M, Shokralla S, Zhou X, Singer GAC, Baird DJ: **Environmental barcoding:**
19 **a next-generation sequencing approach for biomonitoring applications using**
20 **river benthos.** *PloS one* 2011, **6**(4):e17497.
21
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
25 16. Bienert F, Dedanieli S, Miquel C, Coissac E, Poillot C, Brun JJ, Taberlet P: **Tracking**
26 **earthworm communities from soil DNA.** *Molecular Ecology* 2012, **21**(8):2017-2030.
27
28 17. Baird DJ, Hajibabaei M: **Biomonitoring 2.0: a new paradigm in ecosystem**
29 **assessment made possible by next-generation DNA sequencing..** *Molecular*
30 *Ecology* 2012, **21**(8):2039-2044.
31
32 18. Gillett CP, Crampton-Platt A, Timmermans MJ, Jordal B, Emerson BC, Vogler AP: **Bulk**
33 **de novo mitogenome assembly from pooled total DNA elucidates the phylogeny**
34 **of weevils (Coleoptera: Curculionoidea).** *Molecular biology and evolution*
35 2014:msu154.
36
37 19. Krulwich R, LANDER E: **Cracking the Code of Life.** In.: Public Broadcasting Service;
38 2001.
39
40 20. Shokralla S, Gibson JF, Nikbakht H, Janzen DH, Hallwachs W, Hajibabaei M: **Next-**
41 **generation DNA barcoding: using next-generation sequencing to enhance and**
42 **accelerate DNA barcode capture from single specimens.** *Molecular Ecology*
43 *Resources* 2014, **14**(5):892-901.
44
45 21. Shokralla S, Porter TM, Gibson JF, Dobosz R, Janzen DH, Hallwachs W, Golding GB,
46 Hajibabaei M: **Massively parallel multiplex DNA sequencing for specimen**
47 **identification using an Illumina MiSeq platform.** *Scientific reports* 2015, **5**.
48
49 22. Meier R, Wong W, Srivathsan A, Foo M: **\$1 DNA barcodes for reconstructing**
50 **complex phenomes and finding rare species in specimen - rich samples.**
51 *Cladistics* 2016, **32**(1):100-110.
52
53
54
55
56
57
58
59
60
61
62
63
64
65

- 1
2
3
4
5
6
7
8
9
10
11
12
13
14
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
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.
 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.
 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.
 26. 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.
 27. Kumar S, Stecher G, Tamura K: **MEGA7: Molecular Evolutionary Genetics Analysis version 7.0 for bigger datasets.** *Molecular biology and evolution* 2016, **33**(7):1870-1874.
 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.
 29. Li H, Durbin R: **Fast and accurate short read alignment with Burrows–Wheeler transform.** *Bioinformatics* 2009, **25**(14):1754-1760.
 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.
 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. 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.
 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.
 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.
 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.
 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.

1
2
3
4
5
6
7
8
9
10
11
12
13
14
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

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 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.

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

Figure 4. Discrepancies between Sanger and HIFI barcodes in the 1st plate, showing potential heterozygotes (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

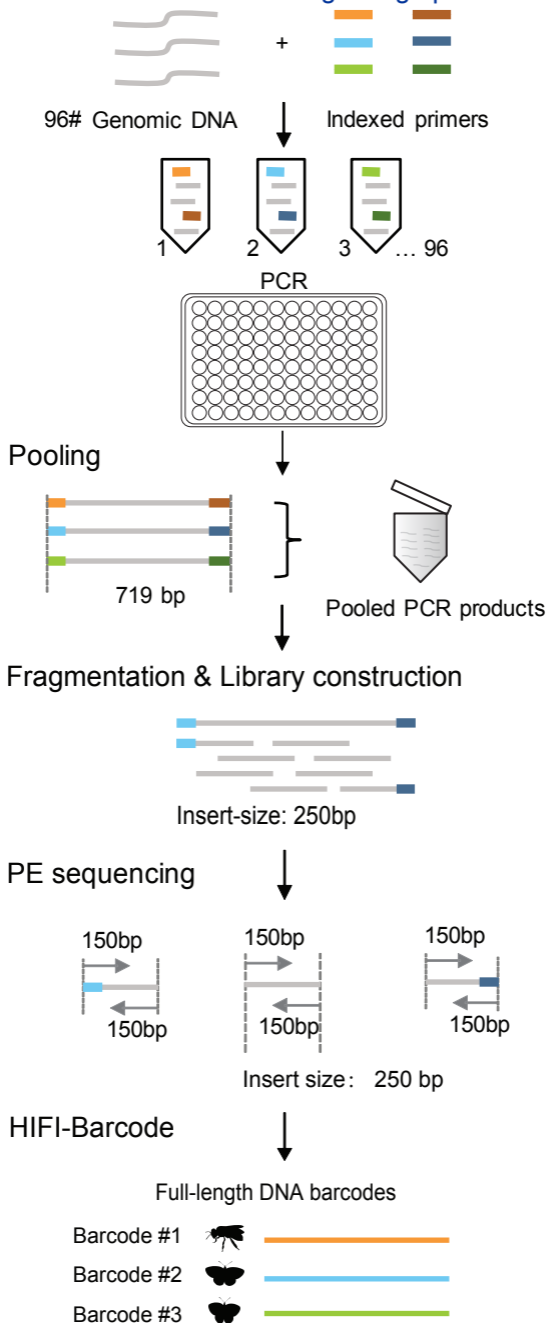
Supplemental Table S4. Non-target sequences detected by HIFI-Barcode

Supplemental File S1. Results of HIFI-barcode

Figure 1

Indexed PCR

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



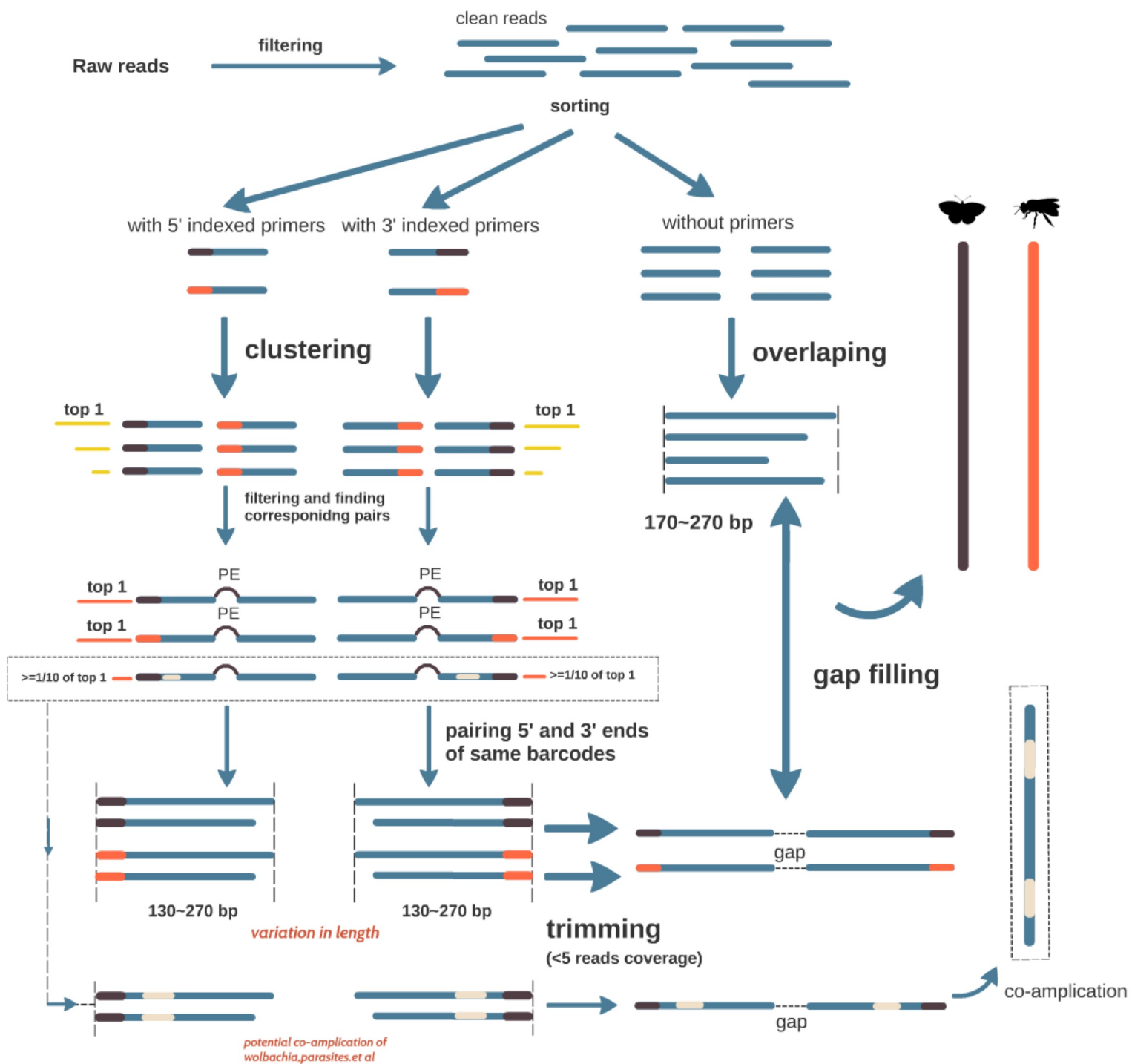
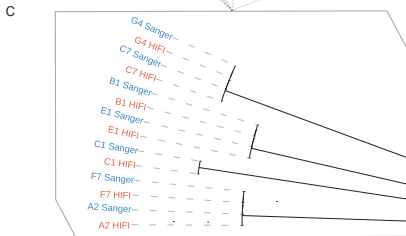
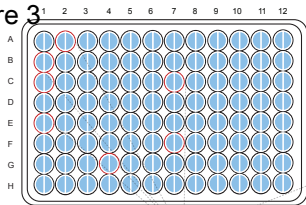
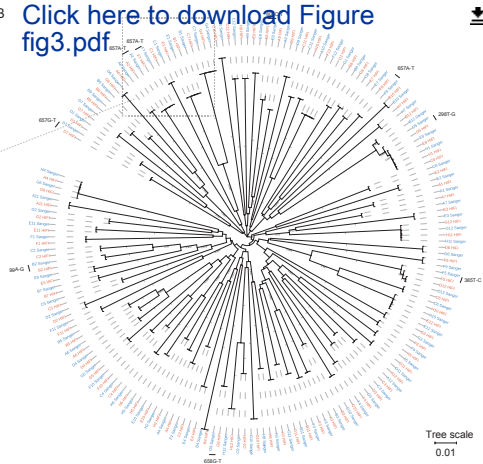


Figure 3

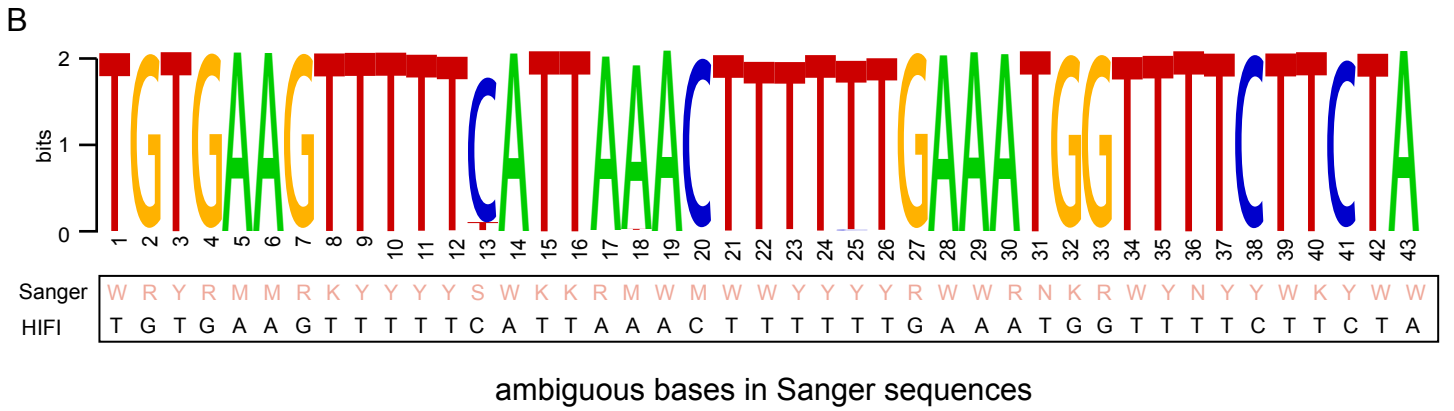
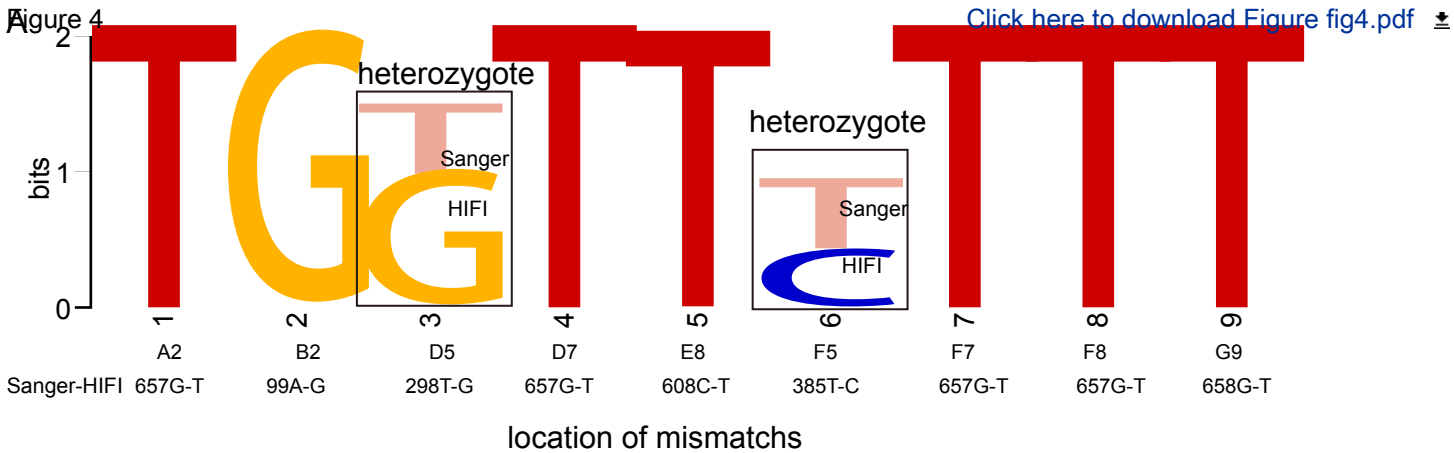


B [Click here to download Figure fig3.pdf](#)



D

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









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





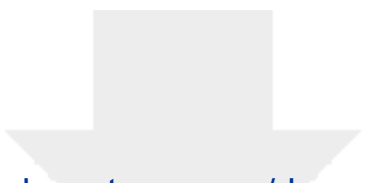
Click here to access/download
Supplementary Material
supplemental table S4.xlsx






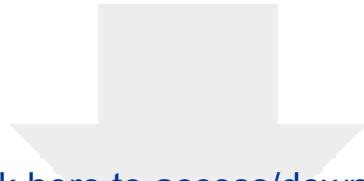
Click here to access/download
Supplementary Material
supplementary figure S1.png





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

