Insect science

Insect Science (2021) 28, 127–143, DOI 10.1111/1744-7917.12749

ORIGINAL ARTICLE

Plant–herbivorous insect networks: who is eating what revealed by long barcodes using high-throughput sequencing and Trinity assembly

Xiao-Man Zhang[1], Zhi-Yong Shi[1], Shao-Qian Zhang[2], Peng Zhang[2], John-James Wilson[3,4], Chungkun Shih[1,5], Jing Li[1], Xue-Dong Li[1], Guo-Yue Yu[6] and Ai-Bing Zhang[1]

[1]*College of Life Sciences, Capital Normal University, Beijing, China;* [2]*School of Life Sciences, Sun Yat-sen University, Guangzhou, China;* [3]*Vertebrate Zoology at World Museum, National Museums Liverpool, Liverpool, United Kingdom;* [4]*Department of Microbiology and Parasitology, Faculty of Medical Science, Naresuan University, Phitsanulok, Thailand;* [5]*Department of Paleobiology, National Museum of Natural History, Smithsonian Institution, Washington, DC, USA and* [6]*Institute of Plant and Environment Protection, Beijing Academy of Agriculture and Forestry Sciences, Beijing, China*

Abstract Interactions between plants and insects are among the most important life functions for all organism at a particular natural community. Usually a large number of samples are required to identify insect diets in food web studies. Previously, Sanger sequencing and next generation sequencing (NGS) with short DNA barcodes were used, resulting in low species-level identification; meanwhile the costs of Sanger sequencing are expensive for metabarcoding together with more samples. Here, we present a fast and effective sequencing strategy to identify larvae of Lepidoptera and their diets at the same time without increasing the cost on Illumina platform in a single HiSeq run, with long-multiplex-metabarcoding (*COI* for insects, *rbcL*, *matK*, *ITS* and *trnL* for plants) obtained by Trinity assembly (SHMMT). Meanwhile, Sanger sequencing (for single individuals) and NGS (for polyphagous) were used to verify the reliability of the SHMMT approach. Furthermore, we show that SHMMT approach is fast and reliable, with most high-quality sequences of five DNA barcodes of 63 larvae individuals (54 species) recovered (full length of 100% of the *COI* gene and 98.3% of plant DNA barcodes) using Trinity assembly (up-sized to 1015 bp). For larvae diets identification, 95% are reliable; the other 5% failed because their guts were empty. The diets identified by SHMMT approach are 100% consistent with the host plants that the larvae were feeding on during our collection. Our study demonstrates that SHMMT approach is reliable and cost-effective for insect-plants network studies. This will facilitate insect-host plant studies that generally contain a huge number of samples.

Key words high-throughput sequencing; host plants; insect network; metabarcoding; phytophagous insects

Introduction

Interactions between insects and their host plants have been an important focus of evolutionary and ecological studies. The evolutionary history of plants and their associated insects, such as lepidopterans and their hosts (Powell *et al.*, 1998), are intertwined (Futuyma & Agrawal, 2009). Therefore, it is particularly important to find an effective method to establish the association between them.

Lepidoptera is one of the largest groups of herbivorous insects worldwide (Donahue, 1995), and caterpillars feed on almost all the orders of angiosperms and gymnosperms (Powell *et al.*, 1998), and many are major pests in

Correspondence: Ai-Bing Zhang, College of Life Sciences, Capital Normal University, Beijing, China. Tel: +86 10 68901860; email: zhangab2008@mail.cnu.edu.cn

agriculture and forestry (Fitt, 1989; Talekar & Shelton, 1993). However, associating caterpillars and their host plants has proven challenging. Larvae rearing has been an important, but time-consuming method to determine the host plant relationship, for example, Bainbridge-Fletcher (Fletcher, 1920) reared 650 species over 20 years to establish host plants of insects in India. Besides rearing, morphological analyses of gut contents and feces, protein electrophoresis (Traugott, 2003) and stable isotope analysis of herbivorous animal tissues (Ponsard & Arditi, 2000) has been used in animal diet studies. However, these methods require experts to identify the plant material. Erroneous records of Lepidoptera host plants are inevitably found in, otherwise erudite, early work (Sattler, 1967).

DNA barcoding was proposed as a relatively simple and quick tool for species identification (Hebert Cywinska *et al.*, 2003; Hajibabaei *et al.*, 2006). The technique was subsequently applied to the analysis of animal diets through DNA barcoding gut contents or fecal samples (King *et al.*, 2008; Valentini *et al.*, 2009; Carreon-Martínez & Heath, 2010; Lim *et al.*, 2017), facilitating many ecological studies (Valentini *et al.*, 2009). The process includes DNA extraction, selection of appropriate DNA barcode gene regions, polymerase chain reaction (PCR) amplification and sequencing (Wilson, 2012) and bioinformatics (Wilson *et al.*, 2018). For investigations of plant-animal trophic interactions, *rbcL* is a commonly sequenced gene region for plant identification (Hoss *et al.*, 1992; Kajtoch *et al.*, 2015). Short *ITS*, *trnL* and P6 loop fragments (usually shorter than 300 bp) are also frequently used when analyzing gut contents or fecal samples because the plant tissue has been digested and DNA is likely to be highly degraded (Matheson *et al.*, 2008; García-Robledo *et al.*, 2013). However, it has been reported that short sequences often do not have sufficient resolution to distinguish plant species (Little, 2014), inevitably reducing the taxonomic precision of studies relying on molecular approaches.

Full-length plant DNA barcodes (about 700 bp for *rbcL* and 1000 bp for *matK*) have been extracted from the guts of insects via Sanger sequencing (Kitson *et al.*, 2013; Kajtoch, 2014; Kajtoch *et al.*, 2015). However, Sanger sequencing can only be applied in situations where only one host plant species is likely to be found in the gut contents or feces of the herbivore (Lim *et al.*, 2017), which is not applicable to generalist insects, which likely have a mix of host plant DNA in their guts. Cloning is a useful method to enable dietary analysis of polyphagous animals (Jo *et al.*, 2016) but it is labor-intensive (Zeale *et al.*, 2011). Furthermore, the cost increases substantially as the number of samples requiring Sanger sequencing increases (Vesterinen *et al.*, 2013; Yang *et al.*, 2016). High-throughput sequencing, with multiplex capabilities, could analyze bulk samples (Zhou *et al.*, 2013) and solve the problem of sequencing the diets of omnivorous animals (Vesterinen *et al.*, 2013). The Illumina platform has been used widely for omnivorous animal diet sequencing but with some limitations in obtaining longer barcodes (no longer than 600 bp) (Kartzinel *et al.*, 2015); however, universal barcodes are usually longer than 600 bp (Kress *et al.*, 2005). Long amplicons could directly sequence using PacBio Sequel Systems, but the error rate of PacBio sequencing was higher than that of next generation sequencing (NGS) when the fragment was shorter than 800 bp, and it is more suitable for 1–6 kb inserted fragments; however, the length of several universal barcodes are from 600 bp to 1000 bp. For MinION, Oxford Nanopore's third-generation sequencing, it has been reported that as an exciting step in a new direction, it need dramatic decreases in error rates (Mikheyev & Tin, 2014). Although the accuracy has been improved, Long-read PacBio and Nanopore sequencing are more suitable for whole-genome *de novo* assembly, not for 600–1000 bp amplicons.

Here we present a novel method for studying plant–insect interactions. The method (SHMMT) uses parallel tagged amplicon sequencing (PTAS) of insect (*COI*) and plant DNA barcodes (*rbcL*+*matK*+*ITS*+*trnL*) on an Illumina platform and sequence assemble by Trinity which could obtain full long-read.

Materials and methods

Barcoding database of plant

A database of DNA barcodes (*trnL*, *rbcL*, *matK*, *trnH-psbA*, *ITS*) (Table S1) for plant species found in Beijing was established in 2016. The plants were identified by morphology and five DNA barcodes. An algorithm combining *rbcL*+*matK*+*ITS*+*trnL* which had the highest discrimination success (100% for family and genus level, 55% for species level) (Fig. S1) was used, and all plants in the database were identified to species level by morphology and a molecular, phylogenetic tree of *rbcL* gene reconstruction was used to verify the accuracy of the identification (Fig. S2).

Insect samples, DNA extraction, and PCR

Five caterpillars fed on known different plants in our laboratory were used as positive control. Live caterpillars of *Ampelophaga rubiginosa* (from Shijiazhuang, Hebei Province), *Dendrolimus punctatus* (from Dawu, Hubei Province), *Gastropacha populifolia* (from

Labagoumen, Beijing), and *Helicoverpa armigera* (from Chinese Academy of Agricultural Science culture), were maintained in our laboratory and they were only fed on *Vitis vinifera*, *Pinus massoniana*, *Salix matsudana* and *Nicotiana tabacum*, respectively. Control caterpillars were subjected to different starvation times to determine the effects of starvation on larval diet identification (*A. rubiginosa* is fresh feeding control, *H. armigera* frozen at −80 °C for 1 year after feeding on *N. tabacum*, *D. punctatus* starved for 24 h after feeding on *P. massoniana*, *G. populifolia* starved for 24 h and 48 h, respectively, after feeding on *S. matsudana*). A total of 58 caterpillar samples (details in article appendix) were collected away from host plants in Beijing. The larvae were immediately preserved in absolute ethanol and frozen at −20 °C the same day.

To reduce the concentration of insect DNA relative to plant DNA, in most cases, the guts of larvae were used as the DNA extraction material (Krehenwinkel *et al.*, 2017). However, for small larvae with a body length of less than 1 cm or width of less than 0.2 cm, the whole animal was used. DNA material from the guts or whole larvae was isolated using a FastDNA™ Spin Kit for Soil Kit (Mpbio), which has the advantages of short extraction time, but a high DNA purity and recovery rate (Ariefdjohan *et al.*, 2010). A Nanodrop (Thermo Scientific™) was used to test the DNA concentration and purity.

Five DNA barcodes were used in this study, *COI* (Lep F and Lep R) for insect identification, and four DNA barcoding regions for plants, *rbcL* (F1 and 724R), *trnL* (A49325 [c] and B49863 [d]), *matK* (390F and 1326R) and *ITS* (ITS5a and ITS4) (Table S2) were used for diet identification. In addition, *trnH-psbA* (psbA3′f and trnH2) was used to establish a plant database but not for diet identification. The DNA of the 63 insect samples used in this study were subjected to individual PCR, and sequencing by two strategies, namely Sanger and NGS (Illumina MiSeq PE300) (S + NGS), and only HTS (high-throughput sequencing, Illumina HiSeq X Ten) (SHMMT approach), while S + NGS was used to verify the accuracy of the SHMMT approach. Because of the higher cost of method 1, only two regions (*TrnL* and *rbcL*) were used for plant identification in method 1 and four regions in method 2.

Method 1: Sequencing by Sanger and NGS

The PCR products were individually sequenced using Sanger technology for 60 insect samples (only one individual was collected at the collection site of the 60 insect samples and gut contents usually contain DNA from only one host plant of the caterpillar) while NGS was used for the remaining three samples (more individuals we collected of these three species and at least three individuals were used from each sample; they are not monophagous) (Fig. S3). The quality of Sanger sequences was checked using SeqMan of DNA STAR LASERGENE. v7.1. The specific steps followed are shown in the article appendix.

Method 2: SHMMT approach

The five target DNA barcode regions were amplified separately, and the products were purified using a Purification Kit (Ensure Biologicals, Shanghai, Art. No: EK03351) according to the manufacturer's protocols. The total quantity of each purified PCR product was not less than 2 μg and the concentration was measured by Nanodrop (each product was greater than 20 ng/μL). The PCR products of five DNA barcodes were pooled by each sample using an equal volume. In addition, HHYP (No.43, No.59 were mixed) as one mixed sample including 17 individuals (14 species), was supposed to be one polyphagous insect that ate seven plant species, and was used to test the accuracy of SHMMT approach for polyphagous insect identification; the total concentration of HHYP was consistent with other samples. A total of 64 samples were used in the SHMMT approach.

Uniquely indexed libraries were prepared for all samples (63 individual insects and one pooled sample - HHYP) according to Feng *et al.* (2016), with slight modifications. Three hundred nanograms of DNA of each purified pool sample was used and randomly sheared by double-stranded DNA fragments in 20 μL reactions; the fragments were then blunt-end-repaired and A-tails added at the 3′ end. After that, unique indexes were ligated to fragments via TA-ligation.

All 64 libraries (index to each libraries) were combined. Suitable fragments (200–500 bp) of the final pool were selected on an agarose gel. The final DNA library was sent to ANOROAD (Beijing, China) and a sequencing library was constructed using the TruSeq DNA Sample Preparation kit for sequencing on an Illumina HiSeq X Ten. The raw reads were processed by ANOROAD first, library barcodes were removed by ANOROAD, which was added to distinguish different libraries sequencing in the same lane, and low-quality data were filtered. This resulted in 22 G of raw data.

Sequencing paired-end reads were sorted based on index, and the linkers were removed during sorting (Feng *et al.*, 2016). The sorted paired-end reads of each 64 species were assembled using Trinity (Grabherr *et al.*, 2011), as per the default parameters. The depth of each

assembled contig was calculated using BOWTIE (Langmead *et al.*, 2009) and SAMTOOLS (Li *et al.*, 2009), Python script written specially for this study was used to assemble all the samples, calculate the depth of each contig, add the sample number to each contig and incorporate all contigs into one text; we finally got a text that includes all the contigs of all samples. In order to obtain high-quality contigs, contigs with an average sequencing depth $>1000\times$ (based on positive controls, the summary of the OUT above $10\times$ detected with depth for five positive controls samples is showed in Table S3. OUTs at depth $<100\times$ are more likely by assemblies mistake, OUT at $100\times <$ depth $< 1000\times$ are more like by contamination) and sequence length within the expected range (the lengths of *COI*, *ITS* and *rbcL* of 500–900 bp, *matK* of 800–1100 bp, *trnL* of 400–800 bp) and cut off 20 bp at both ends (primer length is ~20 bp) were retained for further analysis. For the sample of HHYP, a slightly less strict filter was applied: contigs with a sequencing depth $>300\times$ and length >270 bp were retained.

Genetic designation was added (*COI*, *trnL*, *rbcL*, *matK* and *ITS*) based on the local reference database by Blastn, and sorted by Python script written specially for this study. The sorted genes were first blast in National Center for Biotechnology Information (NCBI), only the cover >80% and ident >98% (based on GenBank blast) were retained by a Python script written specially for this study based on the downloaded XML text. Ninety-eight percent of identified thresholds were used in other plant barcodes studies (Valentini *et al.*, 2009; Heise *et al.*, 2015). This step was used to remove poor assembled sequences and non-flora and fauna sequences, such as fungi, and five texts were finally obtained for further analysis (Fig. 1). The process for our bioinformation analysis has a home page: https://github.com/zhangab2008/HSMMT_project.git.

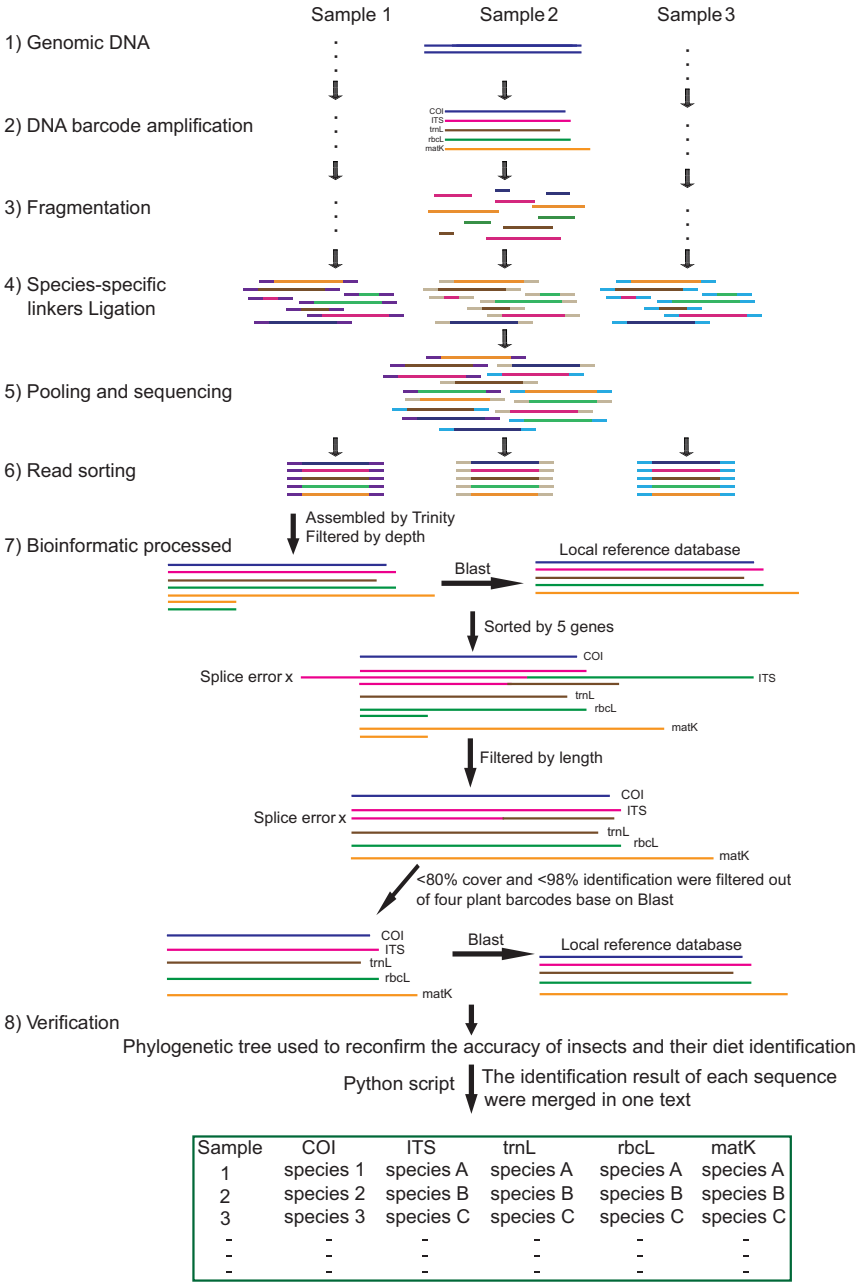
Data analysis

All the sequences were first compared with the local reference database (*trnL*, *rbcL*, *matK* and *ITS* for plants, *COI* for insect species, information see Table S1) by NCBI Blastn; only >99% hits (*trnL*, *rbcL* and *matK* gene) or >98% hits (*COI* and *ITS* gene; because *ITS* shows more intraspecific variations than the other three plant DNA barcoding loci) and >99% coverage were accepted for identifications at the species level. It has been reported that 99% identification threshold of *trnL* and *rbcL* were used in another study (Kajtoch *et al.*, 2015); *matK* and *rbcL* were usually considered as genus-level identification barcodes, so 99% threshold was selected. *COI* shows different intraspecific variations among the species of Lepidoptera (Hajibabaei *et al.*, 2006), so 98% threshold was selected. All the sequences were checked again by NCBI blast against GenBank, using the same parameters (but 80% coverage), to identify any species that were not included in the local reference database. For food plant identification, based on the result of the positive control, only species matches with at least two DNA barcoding regions were accepted; this can lead to more reliable results. For insect identification, *COI* genes >95% (overall mean distance in genus is greater than 5% and the maximum value of pairwise distance in genus is greater than the difference value of identification rate of this sample) hit was scored as identification at the genus level, while if overall mean distance in family and the maximum value of pairwise distance in family were greater than the difference value of identification rate of the sample, it was accepted as identification at the family level.

The PCR and sequencing success, as well as the rate of species identification between method 1 and the SHMMT approach were compared (Fig. 2). The exact length of assembled contig obtained by SHMMT approach are shown in Table 1, where different symbols represent different situations during the experiment.

Phylogenetic tree was used as an alternative approach to test the reliability of assembled contig in the SHMMT approach. *COI* gene and *rbcL* gene from method 1 (62 individuals of *COI* and 56 individuals of *rbcL*) and method 2 (57 individuals of *COI* and 57 individuals of *rbcL*) respectively were combined to reconstruct the phylogenetic tree. *rbcL* gene was used to reconstruct a tree of the host plant, because the *rbcL* gene can be easily amplified and aligned than the other three genes (*trnL*, *matK*, *ITS*). The larvae of Chrysomelidae (two species) and Tenthredinidae (three species, four individuals) were used as the out group of *COI* gene, while *Pinus tabuliformis* was used as the out group of *rbcL* gene. Both *COI* gene and *rbcL* gene were aligned respectively using MEGA 6.0 (Tamura *et al.*, 2013). A maximum-likelihood tree was reconstructed (RAxML). The best-fitting nucleotide substitution model test was obtained using AIC in JMODELTEST 2.1.7 (Posada, 2008); 1000 bootstraps were used to assess the node support by RAxML. The tree was visualized and edited by iTOL (http://itol.embl.de/itol.cgi) (Fig. S4A).

The phylogenetic trees of *COI* and *rbcL* were linked based on the interactions of larvae-host plants (Fig. S4A). The lines in different colors represent different states of the larvae when they were collected (Fig. S4A and B), and the match ratio of the diet was calculated based on molecular result (Fig. S5). *Bipartite* package in *R* (Dormann *et al.*, 2008) was used to construct the network of larvae and their host plants (Fig. S4C). The diet results



Sample	COI	ITS	trnL	rbcL	matK
1	species 1	species A	species A	species A	species A
2	species 2	species B	species B	species B	species B
3	species 3	species C	species C	species C	species C
-	-	-	-	-	-
-	-	-	-	-	-
-	-	-	-	-	-

Fig. 1 Schematic overview of sequencing in a single Hiseq run and long-multiplex-metabarcoding obtained by Trinity (SHMMT) approach and bioinformatic analysis pipe line. 1) DNA extracted and barcodes selected for identification of insects and host plants. 2) Barcodes of different samples amplified, different markers produced from polymerase chain reaction for the same sample are pooled and purified. 3) DNA fragmentation. 4) Adapter ligation was added to both ends of the fragments. 5) Fragment sizes were selected (250 ~ 500 bp) by gel electrophoresis, different libraries were pooled, final size selection and sequencing was achieved by Illumina HiseqX Ten. 6) Reads were sorted by index. 7) Sorted reads were assembled by *de novo* Trinity into full-length contig (finally, full-length sequence of universal barcodes was obtained), depth <1000 were discarded. The assembled contigs were blast against local reference database and five genes sorted by Python script written specifically for this study. Lengths were sorted based on the length of each barcode. Those with <80% cover and <98% identity were discarded based on blast against GenBank (*.xml document). All the high-quality genes were blast against local reference database again. 8) Phylogenetic tree was used to reconfirm the accuracy of identification of SHMMT approach. The identity result of each barcode was merged into one Excel file by Python script written specifically for this study.

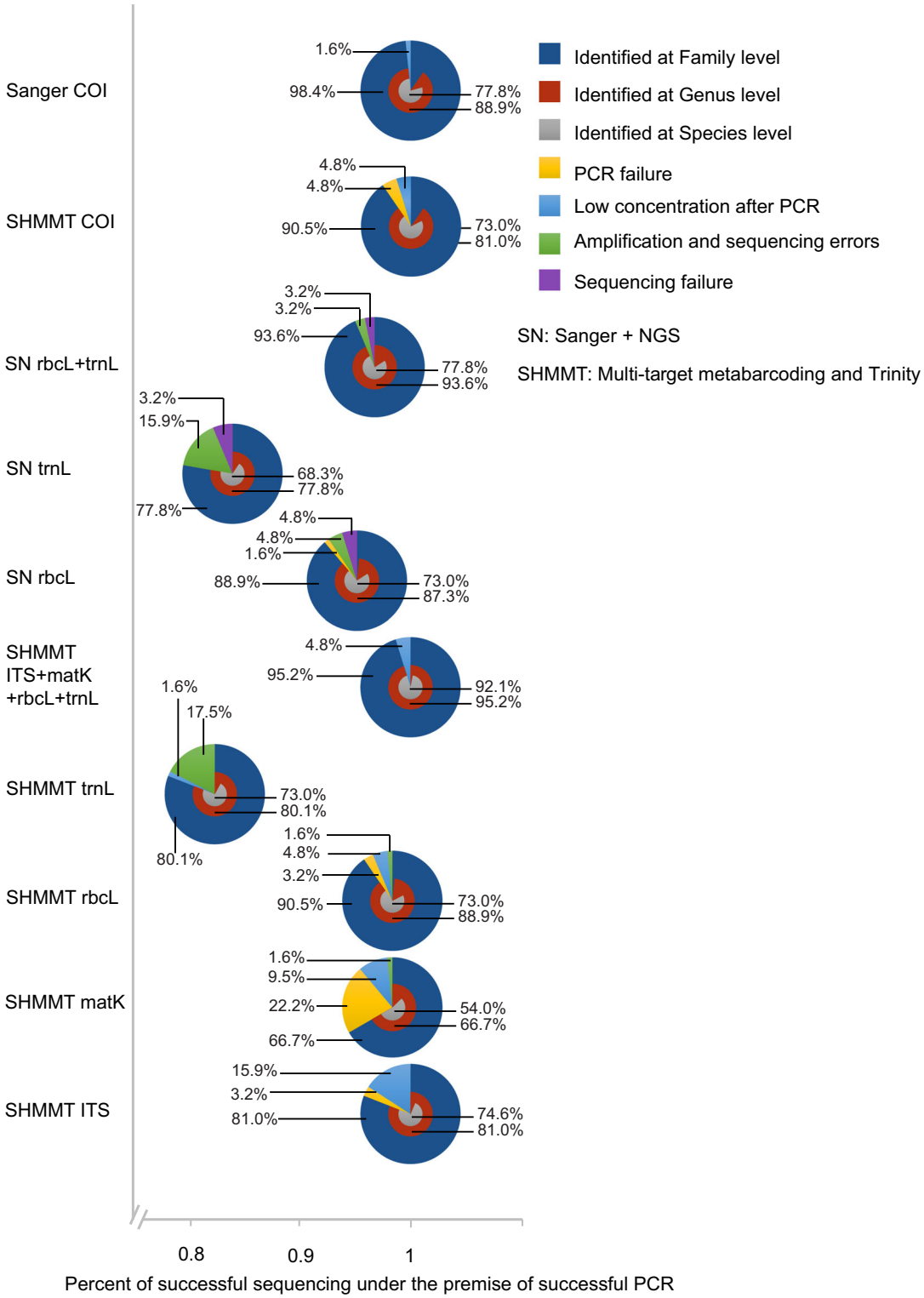


Fig. 2 Comparison of the performance of each barcoding locus by Sanger + next generation sequencing (S + NGS) and sequencing in a single Hiseq run and long-multiplex-metabarcoding obtained by Trinity (SHMMT). Two barcoding loci were used in method 1 (S + NGS) and five barcoding loci barcodes were used in method 2 (SHMMT). Dark blue, red and gray represent the percentage of discrimination at different taxonomic levels for single-locus barcodes or combination barcodes separately.

Table 1 Efficiency of amplification, sequencing and assembling of sequencing in a single Hiseq run and long-multiplex-metabarcoding obtained by Trinity (SHMMT) approach.

Number	ID	Insect species	*COI* (bp length)	Diet (species)	*ITS* (bp length)	*rbcL* (bp length)	*matK* (bp length)	*trnL* (bp length)
MLC	MKY20160121	*Helicoverpa armigera*	√ 724	*Nicotiana tabacum*	※781	√ 746	√ 969	√ 580
LBGM1	LBGM20160910	*Gastropacha populifolia*	√ 749	*Salix matsudana*	√ 752	√ 757	√ 982	※762
LBGM6	LBGM20160912	*Gastropacha populifolia*	√ 704	*Salix matsudana*	LC_X 806	√ 690	○	○
HB	HB20160910	*Dendrolimus punctatus*	√ 727	*Pinus massoniana*	LC_X	√ 761	○	√ 553
1	SSD20160425.01	*Malacosoma neustria*	√ 711	*Populus tomentosa*	√ 734	√ 738	√ 966	√ 664
2	SSD20160425.02	*Micromelalopha sieversi*	√ 728	—	○	LC_X 775	○	X 749
4	JG20160522.02	*Tomostethus*	√ 746	*Fraxinus rhynchophylla*	√ 790	√ 757	√ 989	√ 588
5	JG20160522.03	*Tomostethus*	√ 723	*Fraxinus rhynchophylla*	√ 826	√ 778	√ 971	√ 622
6	JG20160522.04	*Apareophora*	√ 719	*Fraxinus rhynchophylla*	√ 780	√ 807	√ 946	√ 554
7	JG20160522.05	Chrysomelidae	√ 757	*Fraxinus rhynchophylla*	√ 744	√ 769	√ 1015	√ 620
				Quercus dentata	√ 720	Ø 280	‖	‖
8	SS20160601.01	*Naxa seriaria*	√ 711	*Deutzia corymbosa*	※707	√ 737	√ 957	√ 636
9	SS20160601.02	*Calyptra lata*	√ 718	*Menispermum dauricum*	√ 772	√ 774	√ 1008	√ 370
10	SS20160601.05	*Chrysomela populi*	√ 763	*Populus nigra*	√ 743	√ 774	√ 982	√ 491
11	SS20160601.04	*Xylena formosa*	√ 742	*Dioscorea nipponica*	LC_X 762	√ 769	√ 960	※666
13	SS20160601.06	*Agenocimbex maculatus*	LC_X	*Celtis sinensis*	√ 732	√ 769	√ 977	√ 563
14	SS20160601.07	*Araschnia levana*	√ 704	*Urtica angustifolia*	√ 743	√ 761	○	※498
15	SS20160601.08	*Artaxa flava*	√ 735	*Deutzia parviflora*	√ 639	√ 769	√ 962	√ 637
16	SS20160601.09	*Euproctis similis*	√ 735	*Ulmus pumila*	√ 740	√ 769	○	X 558
17	SS20160601.10	*Eupsilia transversa*	√ 738	*Ulmus davidiana*	√ 717	√ 742	○	X 649
18	SS20160601.11	*Xylena formosa*	√ 723	*Clematis heracleifolia*	√ 700	√ 752	√ 962	※570
19	SS20160601.12	*Polygonia c-album*	√ 774	*Ulmus davidiana*	√ 750	√ 764	√ 960	√ 642
20	WLS20160623.01	*Euproctis lutea*	√ 734	*Salix caprea*	√ 741	√ 769	√ 986	※431
21	WLS20160623.02	*Cifuna locuples*	√ 712	*Populus koreana*	√ 760	√ 785	√ 959	√ 607
22	WLS20160623.03	*Acronicta hercules*	√ 762	—	LC_X 810	LC_X 772	○	X 562
23	BHS20160716.01	*Helicoverpa armigera*	√ 728	*Brassica pekinensis*	√ 778	√ 777	○	※405
24	BHS20160716.02	*Pieris rapae*	√ 712	*Brassica pekinensis*	√ 739	√ 744	○	X 411
25	JG20160416.01	*Orgyia recens*	√ 721	*Prunus armeniaca*	√ 744	√ 756	LC	√ 614
27	SJZ20160728.01	*Ampelophaga rubiginosa*	○	*Vitis vinifera*	○	√ 762	√ 845	√ 615
28	SS20160805.01	*Callambulyx tatarinovi*	√ 746	—	LC_X 776	LC_X 777	○	X 630
29	SS20160805.02	*Spilosoma meinshanica*	√ 721	*Spiraea trilobata*	√ 784	√ 730	○	Ø 359
30	SS20160805.03	*Macrophya*	LC_X	*Syringa reticulata*	√ 780	√ 746	√ 959	√ 557
31	SS20160805.04	*Paralebeda femorata*	√ 735	*Malus baccata*	√ 740	√ 767	√ 977	√ 406
32	SS20160805.05	*Acanthopsyche nigraplaga*	√ 730	*Syringa reticulata*	※813	○	○	X 584
33	SS20160805.06	Gelechiidae	√ 776	*Ampelopsis humulifolia*	√ 812	√ 792	√ 982	√ 634

(to be continued)

Table 1 continue.

Number	ID	Insect species	*COI* (bp length)	Diet (species)	*ITS* (bp length)	*rbcL* (bp length)	*matK* (bp length)	*trnL* (bp length)
34	SS20160805.07	*Nolathripa lactaria*	√ 602	*Juglans mandshurica*	√ 765	√ 768	√ 976	√ 665
35	SS20160805.08	*Coarica fasciata*	√ 761	*Juglans mandshurica*	√ 820	√ 778	√ 937	√ 626
36	JG20160806.01	*Kentrochrysalis sieversi*	√ 680	*Fraxinus rhynchophylla*	√ 780	√ 742	√ 954	√ 543
37	JG20160806.02	*Sericinus montela*	○	*Aristolochia contorta*	○	○	√ 1012	※Ø 257
38	SS20160807.01	*Narosoideus flavidorsalis*	√ 712	*Fraxinus rhynchophylla*	LC_X 800	√ 787	√ 939	X 614
39	DLS20160807.01	*Porthesia simillis*	√ 763	*Artemisia capillaris*	√ 763	√ 746	√ 943	√ 600
40	DLS20160813.01	*Nephopterix fumella*	√ 706	*Prunus armeniaca*	√ 705	√ 780	√ 960	√ 613
41	BHS20160716.04	*Helicoverpa armigera*	√ 776	*Prunus armeniaca*	√ 763	√ 604	LC_X934	√ 566
42	BHS20160716.05	*Capusa senilis*	√ 726	*Ulmus pumila*	√ 745	√ 736	○	X 619
43	LJZ20160810.01	Noctuidae	√ 747	*Juglans mandshurica*	√ 760	√ 774	√ 959	※ 622
44	SS20160807.02	*Psoricoptera speciosella*	√ 704	*Juglans mandshurica*	√ 781	√ 752	√ 957	√ 617
45	SS20160807.03	*Sciota*	√ 727	*Ulmus davidiana*	√ 736	√ 761	LC_X 958	√ 593
46	SS20160807.04	*Ectropis*	LC_X	*Ulmus pumila*	√ 695	√ 775	LC_X 954	√ 616
47	SS20160807.05	*Choreutis*	√ 731	*Ulmus davidiana*	√ 753	√ 738	√ 970	√ 593
48	SS20160807.06	Crambidae	√ 723	*Juglans mandshurica*	√ 786	√ 737	LC	※426
49	SS20160805.09	*Sphragifera sigillata*	√ 746	*Oryza sativa*	LC_X 795	√ 769	√ 972	√ 647
50	SS20160805.10	*Sphragifera sigillata*	√ 730	*Juglans mandshurica*	√ 782	√ 761	√ 957	√ 594
51	SS20160805.11	*Ophthalmitis albosignaria*	√ 731	*Juglans mandshurica*	√ 805	√ 752	√ 957	√ 617
52	SS20160805.12	*Mamestra brassicae*	√ 734	*Leonurus japonicus*	√ 759	√ 807	√ 955	※576
53	SS20160805.13	Glyphipterigidae	√ 752	*Dioscorea nipponica*	LC_X 797	√ 765	√ 990	√ 652
54	JG20160806.03	*Orgyia recens*	√ 735	*Spiraea trilobata*	√ 730	√ 769	LC	√ 662
55	SS20160805.14	*Ectropis excellens*	○	*Juglans mandshurica*	√ 794	√ 611	√ 950	√ 622
56	SS20160805.15	*Eupithecia*	√ 728	*Philadelphus pekinensis*	√ 709	√ 764	√ 953	√ 602
57	SS20160807.07	*Hydrelia parvulata*	√ 704	*Juglans mandshurica*	√ 766	√ 761	√ 950	√ 594
58	SS20160805.16	*Sphragifera sigillata*	√ 767	*Juglans mandshurica*	√ 795	√ 794	√ 960	√ 659
59	SS20160807.09	*Sphragifera sigillata*	√ 730	*Juglans mandshurica*	√ 792	√ 761	√ 957	√ 617
60	JG20160806.04	Zygaenidae	√ 741	*Rhododendron*	√ 817	√ 798	√ 935	√ 606
61	SS20160807.11	*Hypena squalida*	√ 720	*Ulmus davidiana*	※719	√ 770	LC_X	√ 634
62	SS20160807.12	*Iragaodes nobilis*	√ 735	*Carpinus turczaninowii*	√ 759	LC_X 742	○	※628

(to be continued)

Table 1 continue.

Sample	Inset species	Diet (species)	Concentration rate	*ITS* (bp length)	*rbcL* (bp length)	*matK* (bp length)	*trnL* (bp length)
HHYP (No.43-No.59)	Presume that one larva feed on seven plants in different concentration	*J. mandshurica*	52.9%	√ 819	√ 659	√ 970	√ 652
		U. davidiana	17.6%	Ø 309	√ 695	Ø 355	Ø 389
		O. sativa	5.9%	‖	‖	‖	Ø 473
		L. japonicus	5.9%	Ø 272	‖	√ 973	‖
		D. nipponica	5.9%	‖	Ø 305	Ø 469	‖
		S. trilobata	5.9%	Ø 289	‖	‖	‖
		P. pekinensis	5.9%	Ø 307	‖	Ø 650	Ø 239

√, polymerase chain reaction (PCR), sequencing and assembling successful (full length obtained); ○, PCR failure; ※, error sequences introduced; LC, no result due to the low concentration after PCR; LC_X, error sequence introduced for the low concentration after PCR; Ø, failure to obtain full length (may be due to the low diet concentration in gut); —, no result due to lack of diet in larvae gut; ‖, no result due to relatively low concentration.

of the larvae in this study were also compared with the records in previous studies (Fig. 3).

Results

Comparison of Sanger sequencing and SHMMT of insect identification

COI of 63 insect samples were sequenced by Sanger in method 1; the DNA sequences of 62 insect samples were obtained Table S4 while one insect sample could not be sequenced due to the low concentration of the PCR product. The identification rate of each barcode and the overall mean distance and pairwise distance in genus and family of seven samples (samples that identified to genus and family level) of *COI* are showed in Table S4. For *COI* sequenced in the SHMMT approach, the full length of 100% of the *COI* gene which was successfully sequenced (based on successful PCR and proper concentration of DNA) and recovered by Trinity assembly (Table 1); 4.8% of the *COI* gene failed in PCR amplification and 4.8% failed to be sequenced due to the low concentration of PCR products (Fig. 2).

Comparison of Sanger+NGS and SHMMT of diet identification

In *Sanger+NGS*, diet of 77.7% of samples had species-level matches (only accepting species-level matches at two or more DNA barcoding regions). Results showed that 4.8% of *rbcL* and 15.9% of *trnL* had amplification and sequencing errors due to bias amplification. The sequencing for 4.8% of *rbcL* and 3.2% of *trnL* failed (Fig. 2).

Diet identification (four barcodes: *trnL* + *rbcL* + *matK* + *ITS*) by SHMMT approach The diet identifications of five laboratory-fed insects (four species) using metabarcoding was consistent with what the insects were fed on. There were 98.2% of full length of four plant DNA barcodings recovered (base on successful PCR and proper concentration of DNA) by Trinity assembly (Table 1). Based on the criteria of only accepting species-level matches at two or more DNA barcoding regions, 92.0% of samples were identified to the species level. For each loci, 73.0% of *trnL* (of 63 samples) had species-level matches, 17.5% *trnL* had amplification errors due to amplification bias of *trnL* to contamination sequences, such as bias to *Pinus tabuliformis* and Rhodiola amplification. For *rbcL* gene, 73.0% of *rbcL* had species-level matches, 4.8% low concentration after PCR due to not having

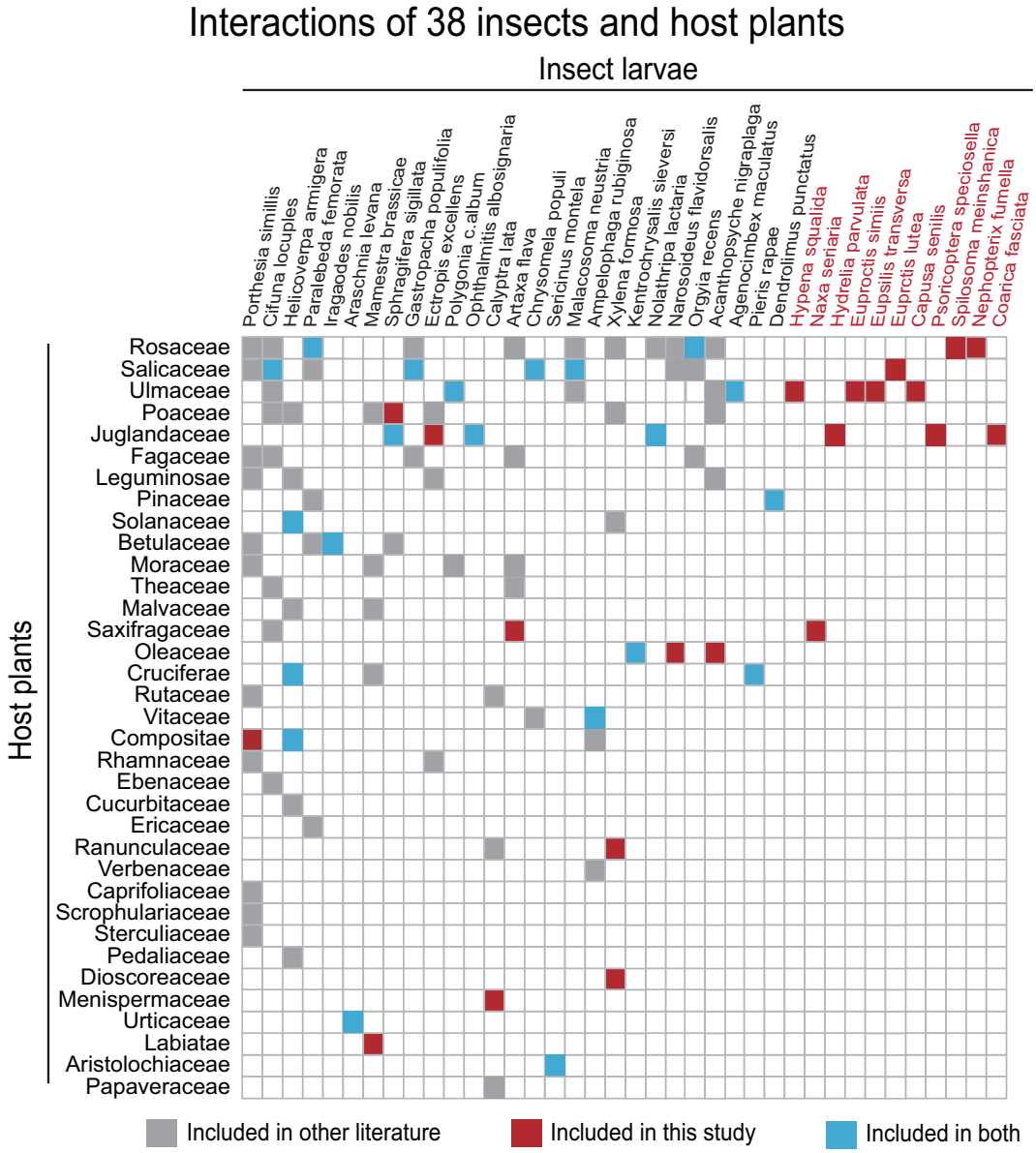


Fig. 3 Reconstruction of network matrix of plant-herbivore larvae that was included in our study and previous studies. Each matrix represents an association between herbivorous larvae (columns) and host plants (rows), interaction shown by red color presents the host plants only included in this study, gray represents host plants not found in this study, blue represents the host plants included in both (this study and previous studies). Species names shown in red color represents host plants with no record in previous study.

yielded high-quality sequences, such as *Spiraea trilobata*, *Deutzia parviflora* and *Alisma gramineum*. *Menispermum dauricum* failed to amplify due to lack species specificity. For *matK* gene, 50.0% had species-level matches, 22.2% had PCR failure due to lack of barcode specificity, such as *Urtica angustifolia*, Ulmus, *Juglans mandshurica*, *Brassica rapa*, *S. trilobata* and *Carpinus turczaninowii*. For *ITS* gene, 74.6% had species-level matches, while 15.9% failed to be sequenced due to the low concentration of PCR products, such as *J. mandshurica* and *M. dauricum*; information detail of each barcodes is shown in Figure 2 and Table S5. For the five positive control samples, the specimen of *Gastropacha populifolia* that was starved for 48 h prior to DNA extraction failed to produce a *trnL* sequence by Sanger+NGS (method 1) or SHMMT approach due to low concentration of PCR products; similarly, the

G. populifolia that was starved for 24 h produced a sequence error of *trnL* (Table S5).

Phylogenetic tree

Sequence of the larvae as well as plants generated from the two sequencing approaches were marked in black and red colors respectively. The phylogenetic tree of *COI* gene and *rbcL* gene showed that 93% sequence of *COI* gene and 95% of *rbcL* gene had no base pair difference from the same species sequenced by both approaches (the bootstrap values were 100 and branch lengths were the same in the phylogenetic tree). For the differential *COI* gene that was obtained by the two sequencing strategies of the same species, there were no more than two base pair different except the *COI* gene of *Chrysomela populi* (five individuals were mixed before DNA extraction, the base pair difference may be due to individual differences). For the *rbcL* gene, 52 of 55 *rbcL* genes had no base difference (Fig. S4A).

Sequencing by Illumina MiSeq in method 1 and method 2

Three insect species (*Helicoverpa armigera*, *Sericinus montela* and another species from Chrysomelidae) were sequenced by Illumina MiSeq in method 1; *rbcL* and *trnL* were used to identify the diet of these three insects, but only one of *trnL* gene (species from Chrysomelidae) obtained full length by Illumina sequencing (the maximum sequencing length of "Illumina MiSeq PE300" was 600 bp and only *trnL* gene of species from Chrysomelidae was shorter than 600 bp). The data of the three species was selected for comparison of the performance of SHMMT approach. The diet of all the three larvae were correctly identified by SHMMT, and 62.5% sequence of metabarcoding of these three species had full length after Trinity assembly (Table S6).

Diet analysis of polyphagous insects in method 2

For HHYP, it was presumed that one larva fed on seven plants to different levels, the total concentration of HHYP (25.5 ng/μL) was consistent with the other samples (about 20–40 ng/μL). Consequently, the concentration of each plant of HHYP were far lower than other samples, therefore, the depth threshold of HHYP (depth $>$300) was set lower than other samples. Full-length genes (*ITS*, *matK*, *rbcL*, *trnL*) of the host plant that was at the highest concentration (it accounted for 52.9% of the total concentration) were obtained. Only *rbcL* gene had the full length when the gene had low concentration at 17.6% (Table 1). The concentrations of the other five plant foods were all 5.9%; only *L. japonicas* of *matK* was obtained with the full length by assembly. Even for the low concentration of mixed foods in one sample, all the seven foods were still identified using SHMMT (Table 1).

Plant–insect network and comparing SHMMT with direct observation

There were 56 interactions between insects and host plants established from the plant–herbivorous insect network (49 interactions among lepidopterans and host plants) (Fig. S4C). The diet of 38 insect species in this study were compared with records from previous studies; the results showed that the diets of monophagous insects identified by metabarcoding were consistent with previous reports (Zhao, 1978; Wu, 2001; Yu, 2015). The diets of oligophagous species, *Polygonia c-album* and *Nolathripa lactarian* identified by metabarcoding (only one host plan was identified in our study) were consistent with plants that were recorded previously (Liu & Wu, 2006). There are 8/16 larvae with polyphagous diets identified in this study which had plant hosts consistent with previous records, while others were different from the previous records and can be considered as supplemental records (Chen, 1999). The food of 11 species named in red had not been reported in previous studies (Fig. 3).

Comparing diet identification using metabarcoding with what was directly observed (33% of larvae were feeding on the host pants when they were collected, 53% were strolling or resting on the plants while 14% were roaming in the field) (Fig. S5) showed that the diet results of the larvae that fed on the host pants identified by metabarcoding were 100% consistent with the directly observed results. For larvae roaming or resting on the plants, 76% of their diets identified by metabarcoding were consistent with the plants they were roaming or resting on; the remaining inconsistencies could be caused by the larvae just passing through the plants when they were collected. The diets of 78% of larvae that were roaming on the field were identified by metabarcoding; other larvae failed to provide the diet result due to hunger (larvae guts were empty when dissected).

Discussion

The objective of this study was to explore a reliable and cost-effective method to test what insects are eating. The SHMMT approach that was developed could identify herbivorous larvae and their host plants simultaneously in a cost-effective way.

Comparing Sanger+NGS sequencing with SHMMT approach

For the identification of insects, it was found that the successful rate of *COI* sequences by Sanger (98.4%, 95% confidence interval: 0.903–0.999) was higher than that by the SHMMT approach (sequencing by Illumina: HiSeq X Ten, which had a 90.5% success rate, 95% confidence interval: 0.798–0.961), because repetitive PCR amplifications and sequencing were performed when the sample had sequencing failure with Sanger, but only one PCR amplification was executed with SHMMT. However, two-sided 95% confidence interval was from −0.015 to 0.174, which includes 0, so there is no statistical significance between the two methods in the successful rate of *COI* sequencing.

For the identification of insect diet (host plant), the SHMMT approach rate of successful sequences (90.5% of *rbcL* and 80.1% of *trnL*) was higher than method 1 (88.9% of *rbcL* and 77.8% of *trnL*). However, two-sided 95% confidence interval was from −0.138 to 0.106 of *rbcL* and two-sided 95% confidence interval was from −0.189 to 0.125 of *trnL* between the two methods, all including 0, so there is no statistical significance between the two methods. Besides considering statistical significance, we also need to consider the reasons for the success and failure of sequencing. Mononucleotide repeats are known to be great challenges in Sanger sequencing and had an effect on assembling of host plant reference library by Sanger in this study. For example, several poly A and poly T regions (more than 10 nucleotides) were detected in *matK* gene (about 1000 bp) of *Malus baccata* and *Aristolochia contorta* respectively, as well as poly C in *ITS* gene of Ulmaceae plants (e.g., *Celtis sinensis*, *Ulmus davidiana*), which reduced the success rate of Sanger sequencing. However, these sequences could easily be generated using SHMMT approach. Another advantage of SHMMT approach is that DNA barcoding loci were randomly interrupted, and the correct result could be acquired by any efficiency segments, which remarkably improve the efficiency of the SHMMT approach. If there are more mononucleotide repeat sequences in one's sample, SHMMT approach will have a significant advantage over method 1.

Metabarcoding was employed in the SHMMT approach to improve the accuracy of the species identification. The diet results identified of *S. montela* via these two sequencing strategies were inconsistent: *Youngia denticulate* was sequenced by Illumina MiSeq PE300 while *Aristolochia contorta* was sequenced by SHMMT. However, it has been recorded that *S. montela* is monophagous and can only feed on *A. contorta* (Wu, 2001), therefore the SHMMT result was more reliable. The main reasons for the error of sequencing by Illumina MiSeq PE300 were contamination (the contamination species could be easily amplified by the bias amplification of different DNA barcodes) and the insufficient DNA barcodes adopted (only two genes, *trnL* and *rbcL* were used in method 1, and the two genes of *A. contorta* are not easily amplified). On the other hand, *matK* loci used in SHMMT approach are proper for *A. contorta* amplification. So the combination of five universal DNA barcoding *COI*+*rbcL*+*matK*+*ITS*+*trnL* by the SHMMT approach was effective in identifying monophagous insects.

The SHMMT approach was not only more accurate but also more economical compared to Sanger+NGS. In method 1, the larvae which fed on two or more plants simultaneously, needed sequencing by NGS (Pompanon *et al.*, 2012). However, with SHMMT approach, whether there are one or more host plants in insect guts, one DNA barcoding or metabarcoding could be sequenced simultaneously in a single HiSeq run, and the full-length gene could be obtained through assembly by Trinity. With the SHMMT strategy, five DNA barcoding regions from 63 samples were sequenced simultaneously in one sequencing run (HiSeq X Ten), at a sequencing cost of typically not more than $250 ($11/G). For the cost of Sanger, DNA extraction: $7.56 × 63 = $476.28; Sanger sequencing: 63 (individual insects) × 5 (four plants barcoding + one insect barcoding) × $5 (bidirectional sequencing of one gene) = ($1575 in total); total cost was ≈ $2015. For the cost of Sanger+NGS (method 1): Sanger: DNA extraction: $7.56 × 60 = $453.6; Sanger sequencing: 60 (individual insects) × 5 (four plants barcoding + one insect barcoding) × $5 (bidirectional sequencing of one gene) = ($1500 in total); total cost was ≈ $2015. NGS: 3 × 43.7 = $131.1 total cost was ≈ $2084. For the cost of SHMMT approach (method 2): DNA extraction: $7.56 × 63 = $476.28; PCR purification: $0.568 × 63 = $35.8; DNA fragmentation: $0.75 × 63 = $47.25; adapter ligation: second set of index was added, 8 × 8 = 64 (number of the samples), so only 8 + 8 = 16 index were needed, and the cost was 16 × $30 = $480; Size Selection: $0.16 × 63 = $10; Illumina sequencing: $250. Total cost ≈ $1299 (note: the costs of adapter ligation synthesis are high, but it can be recycled and re-used repeatedly). For the higher cost of sequencing by Sanger, only two DNA barcoding loci were used for diet identification in method 1.

The control samples (laboratory-fed larvae) revealed several other parameters when using SHMMT method. The diets of freshly fed *Ampelophaga rubiginosa* were identified easily and having good matches to the local reference library. There were error sequences introduced in the *ITS* gene for the *H. armigera* sample (the guts were

frozen for 1 year), and wrong identification results of *ITS* gene of *G. populifolia* (starved for 48 h) and *D. punctatus* (starved for 24 h). For *G. populifolia* (starved for 24 h), an error sequence was introduced in the *trnL* gene. The reasons for these inaccurate host plant sequences could be that the plant DNA had been degraded and digested in the insect gut. When the target DNA is degraded, small amounts of DNA present as contamination may be preferentially amplified during PCR. So at least two accurate full-length plant DNA barcodes were obtained by SHMMT from each laboratory-fed larvae, except in the case of *G. populifolia* starved for 48 h, Therefore, allowing an identification of the host plant identified at species-level under the criteria that at least the identification result of two barcodes were consistent was used in this study.

Long amplicons versus short amplicons

Little (2014) mentioned that the discriminatory power of mini-barcodes noticeably decreased at lower taxonomic levels and the identification rate of best mini-barcode was less than 38.2% at the species level. In our study discriminatory power was 88.9% (95% confidence interval from 0.778 to 0.950) of full-length *rbcL* gene at genus taxonomic levels and 73.0% (95% confidence interval from 0.601 to 0.831) at species level (Fig. 2) which was higher than that of short-*rbcL* gene (230 bases long and 320 bases long) in a previous study (47% identified host plant to genus taxonomic levels) (García-Robledo *et al.*, 2013). Overall the full-length barcode has sufficient resolution to distinguish species.

However, for highly degraded plants in the insect gut, it could be recovered by short barcodes, indicating that the barcodes can be used as alternative barcodes for identification of highly degraded plants. It is possible to identify more foods in the gut by using both short barcodes and long barcodes, but choice of threshold is particularly important especially sequencing depth threshold; depth thresholds should be set for long and short barcodes.

Comparing direct observation with SHMMT approach

Time and cost saving are evident if 90 individuals and five DNA barcodes are used to identify the diet of herbivorous insects; one week is enough to complete the whole laboratory work, such as DNA extraction, PCR, purification and library preparation. Amplicon sequencing by Illumina (by sequencing company) takes no more than half a month, and bioinformatics analyses require 1 to 2 days (on condition that the software and the script have been established). As the sample quantity increases, the cost and time does not increase. Another advantage is reliable, full-length amplicons (up to 1015 bp) followed by SHMMT approach improves the precision of the species identification; 92.1% of diet was identified to the species level. Comparing the metabarcoding identification results with direct observation in this study, metabarcoding showed higher accuracy of identification (Fig. S5), the host plant of the larvae strolling on the road (Fig. S5G) and strolling or resting on the plants (Fig. S5E) could be obtained by metabarcoding, but failed by direct observation. SHMMT approach can be used as a supplement to direct observation, such as the host plant of polyphagous insects and the insects with uncertain host plants (Fig. 3).

Limitation and future direction of SHMMT approach

In this study, it was found that 63 individuals were somewhat inadequate from the economic level of SHMMT approach; for example, in previous study, 320 individuals and five loci were used by PTAS (Puritz *et al.*, 2012). The other limitation of this method was the costs of indexes; the number of indexes were the same as the samples, therefore, for more different samples, more indexes are needed (the costs of species-specific barcode linker synthesis are high, but it can be recycled and re-used repeatedly). This problem has already been resolved by Feng *et al.* (2016), who added the second set of indexes to 3′-G overhang, and cut down the costs. Thus, SHMMT approach is suitable for the diet identification of a great mass of herbivorous insects. Of course, the balance between the cost of the index synthesis and the sequencing should be considered. For example, 20 indexes can test $10 \times 10 = 100$ individuals and index synthesis at a cost of 20 (species-specific barcodes) $\times$ \$30 = \$600, and the cost of HiSeq X Ten sequencing is \$250 (\$11/G), making a total of \$850. If 200 individual samples are to be tested, 15 (species-specific barcodes) $\times$ 15 (species-specific barcodes) = 225 individuals could be tested, $30 \times \$30 = \$900$ for index synthesis and the total cost would be \$900 + \$250 (sequencing cost) = \$1150, but if the 200 individuals are divided into two parts and sequenced twice, only 20 indexes (barcode linkers can be re-used) need to be synthesized and the total costs would be $\$600 + \$250 \times 2 = \$1100$. Therefore, the maximum cost saving depends on the quantity of the samples.

Other problems of this method are contamination during the experiments, DNA barcoding amplification bias and assembly errors. Contamination can be avoided through careful operation during sampling, DNA

extraction, PCR, and so on; for example, by use of sterilized vessels, operating on ultra-clean tables and avoiding cross-contamination. The small amount of contamination can also be avoided to set a threshold in the subsequent biological information step. Meanwhile, this step would eliminate low-concentration diet in the insect's guts, whereas the missing data could be avoided by increasing the number of samples.

For PCR amplification bias, some DNA barcodes are still more aggressive for different plant taxon groups, rare pollution can lead to identification error. For example, *trnH-psbA* marker is not suitable for Zingiberales due to the long repeat A-T (Hollingsworth *et al.*, 2009). This problem can be solved by using metabarcoding; in this study, four plant DNA barcodes were used. Another problem is for the larvae, which feed on more than one host plant, the concentration of the food plants in one gut may be significantly different and may cause a bias amplification; these could easily lead to amplification failure. For example, seven host plants were mixed for the HHYP sample, most of the lower concentration plants fail to get full-length sequences, some even fail to get any segments (Table 1). Based on a previous study (Kajtoch *et al.*, 2015), it was suggested in this study that DNA barcode loci should be amplified separately for each individual (even for the same species) to avoid the amplification bias of unequal concentration. In addition, if two closely related species co-exist in the gut of one insect, it is difficult to identify two species, usually only one plant can be identified (generally the one with the higher concentration of DNA); this may be due to the high similarity of the same gene between the two closely related species, and only one of the sequences was obtained during Trinity assembly. We hope algorithms can be optimized and genes with high similarity can be assembled simultaneously in the future; of course, high precision assembly may introduce some genes with sequencing errors, but we believe that such errors are rare and can be filtered out by sequencing depth.

For assembly problems, once in a while, two barcodes of one sample are spliced together, for example, *ITS* and *COI* genes of No. 54 (*Orgyia recens*) were spliced together, while *ITS* and *rbcL* genes of No. 29 (*Spilosoma meinshanica*) were spliced together. Sometimes two short fragments can be spliced together, causing splicing errors. These can be found easily by original *.depth text (generated by SAMTOOLS). To solve this problem, one should check the depth of the single nucleobase one by one base on *.depth text; if the depth is suddenly reduced in the middle of the sequence, it is generally recognized as a splicing error, and these errors could be rectified manually. These steps are done after the threshold of depth and length, of course one does not need to check the assembled sequence one by one, but only needs to check if the length of the spliced sequencing is equal to the lengths of two DNA barcodes or not. To ensure the accuracy of the assembled contig, blast and construct a phylogeny tree are necessary. Before these, all the sequencing of splicing must be converted into forward sequences. Full-length barcode could be sequenced by PacBio sequencing, no assembly required; however, the costs of PacBio sequencing are much higher than NGS.

The same DNA barcoding loci of several plants were produced after Trinity assembly in one sample called "paralogous gene". If "paralogous gene" was chosen after threshold selection, the larvae were thought to feed on more than one plants. Sometimes the contamination in a previous process would cause a "paralogous gene"; this error can be avoided by careful operation and metabarcoding. There are other *de novo* software for reads assembly, such as SOAP denovo (Li *et al.*, 2010; Xie *et al.*, 2014), and CUFFLINKS (Trapnell *et al.*, 2010).

Conclusion

Full-length universal DNA barcodes of insects and their host plants were applied in this study using the SHMMT approach, incorporating the strategy of metabarcoding, which greatly improved the accuracy of species identification for insects or their host plants. These approaches showed high potential to identify hundreds of insects (monophagy, oligophagy or phytophagous) and their diet DNA using SHMMT approach (Illumina sequencing), which also reduced the costs of sequencing compared with Sanger sequencing. As a result, this method improved the efficiency of diet recognition. To this end, we recommend the combination of five insects and plants universal barcoding *COI*+*rbcL*+*matK*+*ITS*+*trnL* as the metabarcoding algorithm for phytophagous insects and their diet identification by SHMMT approach. For highly digested plant, one can choose long and short barcodes for food identification. The protocol established here is convenient for studying the food web of phytophagous insects. We believe that this strategy can also be applied in the network study of carnivorous insects and omnivorous insects, if suitable metabarcoding target taxon groups are chosen.

Acknowledgments

We thank Prof. Mei-Cai Wei in Central South University of Forestry and Technology for larvae morphological identification. We appreciate all the members in our lab

for larvae collection. This work was supported by the China National Funds for Distinguished Young Scientists (to Zhang, Grant No. 31425023), Natural Science Foundation of China (to Zhang, Grant 31772501), Support Project of High-level Teachers in Beijing Municipal Universities (No. IDHT20180518), Academy for Multidisciplinary Studies, Capital Normal University and also supported by Program for Changjiang Scholars and Innovative Research Team in University (IRT_17R75).

Abbreviations

SHMMT: Sequencing in a single Hiseq run and long-multiplex-metabarcoding obtained by Trinity; HTS: High-throughput sequencing; PTAS: Parallel tagged amplicon sequencing; NGS: Next generation sequencing. S + NGS: Sanger + Next generation sequencing.

Disclosure

The authors declare no conflict of interest.

Data accessibility

DNA sequences (*ITS*, *matK*, *rbcL*, *trnL*) of local reference database of plants: GenBank accessions *ITS*: MG772938–MG772981; *matK*: MG772982–MG773012; *rbcL*: MG-773013–MG773065; *trnL*: MG773066–MG773119), DNA sequences of the larvae: GenBank accession *COI*: MK386578–MK386639. Raw read data of SHMMT approach (method 2) and NGS in method 1 are available from NCBI bioproject PRJNA511943.

References

Ariefdjohan, M.W., Savaiano, D.A. and Nakatsu, C.H. (2010) Comparison of DNA extraction kits for PCR-DGGE analysis of human intestinal microbial communities from fecal specimens. *Nutrition Journal*, 9, 23.

Carreon-Martínez, L. and Heath, D.D. (2010) Revolution in food web analysis and trophic ecology: diet analysis by DNA and stable isotope analysis. *Molecular Ecology*, 19, 25–27.

Chen, Y.X. (1999) *Fauna Sinica Insecta (Vol.16) Lepidoptera: Noctuidae*. Science Press, Beijing. pp. 52–53.

Donahue, J.P. (1995) The Lepidoptera: Form, Function and Diversity. *Annals of The Entomological Society of America*, 88, 590–590.

Dormann, C.F., Gruber, B. and Fründ, J. (2008) Introducing the bipartite package: analysing ecological networks. *R News*, 8, 8–11.

Feng, Y., Liu, Q., Chen, M., Liang, D. and Zhang, P. (2016) Parallel tagged amplicon sequencing of relatively long PCR products using the Illumina HiSeq platform and transcriptome assembly. *Molecular Ecology Resourse*, 16, 91–102.

Fitt, G.P. (1989) The ecology of *Heliothis* species in relation to agroecosystems. *Annual Review of Entomology*, 34, 17–52.

Fletcher, T.B. (1920) Life histories of Indian insects. *Microlepidoptera. Memoirs of the Department of Agriculture in India. Entomlogical Series*, 6, 1–217.

Futuyma, D.J. and Agrawal, A.A. (2009) Macroevolution and the biological diversity of plants and herbivores. *Proceedings of the National Academy of Sciences USA*, 106, 18054–18061.

García-Robledo, C., Erickson, D.L., Staines, C.L., Erwin, T.L. and Kress, W.J. (2013) Tropical plant–herbivore networks: reconstructing species interactions using DNA barcodes. *PLoS ONE*, 8, e52967.

Grabherr, M.G., Haas, B.J., Yassour, M., Levin, J.Z., Thompson, D.A., Amit, I. *et al*. (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology*, 29, 644–652.

Hollingsworth, P.M., Forrest, L.L., Spouge, J.L., Hajibabaei, M., Ratnasingham, S., van der Bank, M. *et al*. (2009) A DNA barcode for land plants. *Proceedings of the National Academy of Sciences USA*, 106, 12794–12797.

Hajibabaei, M., Janzen, D.H., Burns, J.M., Hallwachs, W. and Hebert, P.D. (2006) DNA barcodes distinguish species of tropical Lepidoptera. *Proceedings of the National Academy of Sciences USA*, 103, 968–971.

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

Heise, W., Kubisz, D., Babik, W. and Kajtoch, Ł. (2015) A three-marker DNA barcoding approach for ecological studies of xerothermic plants and herbivorous insects from central Europe. *Botanical Journal of Linnaean Society*, 177, 576–592.

Hoss, M., Kohn, M.H., Paabo, S., Knauer, F. and Schroder, W. (1992) Excremental analysis by PCR. *Nature*, 359, 199–199.

Jo, H., Ventura, M., Vidal, N., Gim, J., Buchaca, T., Barmuta, L.A. *et al*. (2016) Discovering hidden biodiversity: the use of complementary monitoring of fish diet based on DNA barcoding in freshwater ecosystems. *Ecology and Evolution*, 6, 219–232.

Kajtoch, Ł. (2014) A DNA metabarcoding study of a polyphagous beetle dietary diversity: the utility of barcodes and sequencing techniques. *Folia Biologica*, 62, 223–234.

Kajtoch, Ł., Kubisz, D., Heise, W., Mazur, M.A. and Babik, W. (2015) Plant–herbivorous beetle networks: molecular characterization of trophic ecology within a threatened steppic environment. *Molecular Ecology*, 24, 4023–4038.

Kartzinel, T.R., Chen, P.A., Coverdale, T.C., Erickson, D.L., Kress, W.J., Kuzmina, M. *et al*. (2015) DNA metabarcoding illuminates dietary niche partitioning by African large herbivores. *Proceedings of the National Academy of Sciences USA*, 112, 8019–8024.

King, R.A., Read, D.S., Traugott, M. and Symondson, W.O.C. (2008) Molecular analysis of predation: a review of best practice for DNA-based approaches. *Molecular Ecology*, 17, 947–963.

Kitson, J.J., Warren, B.H., Florens, F.B., Baider, C., Strasberg, D. and Emerson, B.C. (2013) Molecular characterization of trophic ecology within an island radiation of insect herbivores (Curculionidae: Entiminae: *Cratopus*). *Molecular Ecology*, 22, 5441–5455.

Krehenwinkel, H., Kennedy, S., Pekár, S. and Gillespie, R.G. (2017) A cost-efficient and simple protocol to enrich prey DNA from extractions of predatory arthropods for large-scale gut content analysis by Illumina sequencing. *Methods in Ecology and Evolution*, 8, 126–134.

Kress, W.J., Wurdack, K.J., Zimmer, E.A., Weigt, L.A. and Janzen, D.H. (2005) Use of DNA barcodes to identify flowering plants. *Proceedings of the National Academy of Sciences USA*, 102, 8369–8374.

Langmead, B., Trapnell, C., Pop, M. and Salzberg, S.L. (2009) Ultrafast and memory efficient alignment of short DNA sequences to the human genome. *Genome Biology*, 10, R25.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N. *et al*. (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25, 2078–2079.

Li, R.Q., Zhu, H.M., Ruan, J., Qian, W.B., Fang, X.D., Shi, Z.B. *et al*. (2010) *De novo* assembly of human genomes with massively parallel short read sequencing. *Genome Research*, 20, 265–272.

Lim, V.C., Clare, E.L., Littlefair, J.E., Ramli, R., Bhassu, S. and Wilson, J.J. (2018) Impact of urbanisation and agriculture on the diet of fruit bats. *Urban Ecosystems*, 21, 61–70.

Little, D.P. (2014) A DNA mini-barcode for land plants. *Molecular Ecology Resourse*, 14, 437–446.

Liu, Y.Q. and Wu, C.S. (2006) *Fauna Sinica Insecta (Vol. 47) Lepidoptera: Lasiocampidae*. Science Press, Bejing. pp. 16–21.

Matheson, C., Muller, G., Junnila, A., Vernon, K., Hausmann, A., Miller, M. *et al*. (2008) A PCR method for detection of plant meals from the guts of insects. *Organisms Diversity & Evolution*, 7, 294–303.

Mikheyev, A.S. and Tin, M.M.Y. (2014) A first look at the oxford nanopore minion sequencer. *Molecular Ecology Resources*, 14, 1097–1102.

Pompanon, F., Deagle, B.E., Symondson, W.O., Brown, D.S., Jarman, S.N. and Taberlet, P. (2012) Who is eating what: diet assessment using next generation sequencing. *Molecular Ecology*, 21, 1931–1950.

Ponsard, S. and Arditi, R. (2000) What can stable isotopes (δ15N and δ13C) tell about the food web of soil macro-invertebrates? *Ecology*, 81, 852–864.

Posada, D. (2008) jModelTest: Phylogenetic Model Averaging. *Molecular Biology Evolution*, 25, 1253–1256.

Powell, J.A., Mitter, C. and Farrell, B. (1998) Evolution of larval food preferences in Lepidoptera. In *Handbook of Zoology*, N.P. Kristensen, Ed. (Volume IV, Arthropoda: Insecta, Part 35). Lepidoptera, Moths and Butterflies, Volume 1: Evolution, Systematics, and Biogeography, Walter de Gruyter, Berlin, New York, pp. 403–422.

Puritz, J.B., Addison, J.A. and Toonen, R.J. (2012) Next-generation phylogeography: a targeted approach for multilocus sequencing of non-model organisms. *PLoS ONE*, 7, e34241.

Sattler, K. (1967) Ethmiidae. *Microlepidaptera Palearctica 2*. (pp. 185). Vienna: Fromme.

Talekar, N.S. and Shelton, A.M. (1993) Biology, ecology, and management of the diamondback moth. *Annual Review of Entomology*, 38, 275–301.

Tamura, K., Stecher, G., Peterson, D., Filipski, A. and Kumar, S. (2013) MEGA6: Molecular Evolutionary Genetics Analysis Version 6.0. *Molecular Biology Evolution*, 30, 2725–2729.

Trapnell, C., Williams, B.A, Pertea, G., Mortazavi, A., Kwan, G., van Baren, M.J. *et al*. (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology*, 28, 511–515.

Traugott, M. (2003) The prey spectrum of larval and adult *Cantharis* species in arable land: An electrophoretic approach. *Pedobiologia*, 47, 161–169.

Valentini, A., Miquel, C., Nawaz, M.A., Bellemain, E., Coissac, E., Pompanon, F. *et al*. (2009) New perspectives in diet analysis based on DNA barcoding and parallel pyrosequencing: The trnL approach. *Molecular Ecology Resourse*, 9, 51–60.

Valentini, A., Pompanon, F. and Taberlet, P. (2009) DNA barcoding for ecologists. *Trends in Ecology & Evolution*, 24, 110–117.

Vesterinen, E.J., Lilley, T.M., Laine, V.N. and Wahlberg, N. (2013) Next generation sequencing of fecal DNA reveals the dietary diversity of the widespread insectivorous predator Daubenton's bat (*Myotis daubentonii*) in Southwestern Finland. *PLoS ONE*, 8, e82168.

Wilson, J.J. (2012) DNA barcodes for insects. *DNA Barcodes: Methods and Protocols* (eds. W.J. Kress & D.L. Erikson), pp. 18–44. Humana Press, New York.

Wilson, J.J., Sing, K.W. and Jaturas, N. (2018) DNA barcoding: Bioinformatics workflows for beginners. *Encyclopedia of Bioinformatics and Computational Biology* (eds. S. Ranganathan, M. Gribskov, K. Nakai & C. Schönbach.), pp. 985–995. Academic Press, Massachusetts.

Wu, C.S. (2001) *Fauna Sinica Insecta (Vol. 25) Lepidoptera: Papilionidae*. Science press, Beijing.

Xie, Y.L., Wu, G.X., Tang, J.B., Luo, R., Patterson, J., Liu, S.L. *et al*. (2014) SOAPdenovo-Trans: *de novo* transcriptome assembly with short RNA-Seq reads. *Bioinformatics*, 30, 1660–1666.

Yang, Y., Zhan, A., Cao, L., Meng, F. and Xu, W. (2016) Selection of a marker gene to construct a reference library for wetland plants, and the application of metabarcoding to analyze the diet of wintering herbivorous waterbirds. *PeerJ*, 4, e2345.

Yu, G.Y. (2015) *Moth in Beijing*. Science Press, Beijing.

Zeale, M.R.K., Butlin, R.K., Barker, G.L.A., Lees, D.C. and Jones, G. (2011) Taxon-specific PCR for DNA barcoding arthropod prey in bat faeces. *Molecular Ecology Resourse*, 11, 236–244.

Zhao, Z.L. (1978) *Economic Insect Fauna of China (vol. 12) Lepidoptera: Lymantridae*. Science Press, Beijing. pp. 44.

Zhou, X., Li, Y.Y., Liu, S.L., Yang, Q., Su, X., Zhou, L.L. *et al*. (2013) Ultra-deep sequencing enables highfdelity recovery of biodiversity for bulk arthropod samples without PCR amplifcation. *Gigascience*, 2, 4.

Manuscript received October 20, 2019
Final version received December 12, 2019
Accepted December 15, 2019

Supporting Information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Table S1 DNA barcoding information of local reference database of plants.

Table S2 The primers of five DNA barcoding loci.

Table S3 The summary of the OTU detected with depth of five positive control samples.

Table S4 The merged identification results of Sanger and next generation sequencing (NGS).

Table S5 The merged identification results of the sequencing in a single Hiseq run and long-multiplex-metabarcoding obtained by Trinity (SHMMT) approach.

Table S6 The performance of plant barcodes (sequencing by next generation sequencing [NGS]) in method 1 and method 2.

Fig. S1 Taxonomic resolution of five plant DNA barcodes (based on blast against GenBank).

Fig. S2 Phylogenetic tree of plants in the local reference database based on *rbcL* gene.

Fig. S3 Schematic overview of Sanger and Illumina MiSeq, sequencing combined Sanger with next generation sequencing (NGS).

Fig. S4 (A): Maximum-likelihood phylogenetic tree reconstructed based on *COI* and *rbcL* genes that were generated by two methods (only bootstraps with a value >50% are presented). (B): The picture of larvae in different status. (C): Plant-herbivore larvae network constructed by "bipartite" package in R.

Fig. S5 Percentage of the diet identification by metabarcoding of larvae in different states.

Article appendix: Sampling sites, DNA barcoding database and detail in method 1.