MOLECULAR ECOLOGY
RESOURCES
Molecular Ecology Resources (2016) **16**, 470–479 doi: 10.1111/1755-0998.12472

Mitochondrial capture enriches mito-DNA 100 fold, enabling PCR-free mitogenomics biodiversity analysis

SHANLIN LIU,*†‡ XIN WANG,*† LIN XIE,† MEIHUA TAN,*† ZHENYU LI,† XU SU,*§ HAO ZHANG,† BERNHARD MISOF,¶ KARL M. KJER,** MIN TANG,*† OLIVER NIEHUIS,¶ HUI JIANG† and XIN ZHOU*†

*China National GeneBank-Shenzhen, BGI-Shenzhen, Shenzhen, Guangdong Province 518083, China, †BGI-Shenzhen, Shenzhen, Guangdong Province 518083, China, ‡Centre for GeoGenetics, Natural History Museum of Denmark, University of Copenhagen, Øster Voldgade 5–7, 1350 Copenhagen, Denmark, §Guizhou provincial Center For Disease Control And Prevention, Guiyang, Guizhou province 550004, China, ¶Zoologisches Forschungsmuseum Alexander Koenig (ZFMK)/Zentrum für Molekulare Biodiversitätsforschung (ZMB), Bonn, Germany, **Department of Entomology and Nematology, UC Davis, Davis, CA 95616, USA

Abstract

Biodiversity analyses based on next-generation sequencing (NGS) platforms have developed by leaps and bounds in recent years. A PCR-free strategy, which can alleviate taxonomic bias, was considered as a promising approach to delivering reliable species compositions of targeted environments. The major impediment of such a method is the lack of appropriate mitochondrial DNA enrichment ways. Because mitochondrial genomes (mitogenomes) make up only a small proportion of total DNA, PCR-free methods will inevitably result in a huge excess of data (>99%). Furthermore, the massive volume of sequence data is highly demanding on computing resources. Here, we present a mitogenome enrichment pipeline via a gene capture chip that was designed by virtue of the mitogenome sequences of the 1000 Insect Transcriptome Evolution project (1KITE, www.1kite.org). A mock sample containing 49 species was used to evaluate the efficiency of the mitogenome capture method. We demonstrate that the proportion of mitochondrial DNA can be increased by approximately 100-fold (from the original 0.47% to 42.52%). Variation in phylogenetic distances of target taxa to the probe set could in principle result in bias in abundance. However, the frequencies of input taxa were largely maintained after capture (R^2 = 0.81). We suggest that our mitogenome capture approach coupled with PCR-free shotgun sequencing could provide ecological researchers an efficient NGS method to deliver reliable biodiversity assessment.

Keywords: biodiversity, gene capture, microarray, mitochondrial genome

Received 29 November 2014; revision received 19 September 2015; accepted 24 September 2015

Introduction

Reliable biodiversity estimates are vital to a multitude of important management decisions concerning the sustainability of ecosystems and exploitation of natural resources (Board 2005; Keesing *et al.* 2010; Naidoo *et al.* 2011). Recently, next-generation sequencing (NGS) has been introduced to estimate biodiversity from mass samples (Porazinska *et al.* 2009, 2010; Hajibabaei *et al.* 2011; Yu *et al.* 2012; Ji *et al.* 2013; Liu *et al.* 2013) or even environmental DNAs (Baird & Hajibabaei 2012; Bienert *et al.* 2012; Coissac *et al.* 2012). The majority of these studies has been based on NGS sequencing of PCR amplicons, which, unfortunately, often leads to taxonomic bias (Taberlet *et al.* 2012; Liu *et al.* 2013). Zhou *et al.* (2013) investigated the feasibility of recovering species richness from homogenized arthropod samples using a whole-genome shotgun approach without amplifying any target genes. This work revealed a strong correlation between read numbers and biomass of sequenced taxa, suggesting a potential metric to estimate relative abundance. In addition, by expanding a single barcode gene to a whole mitogenome, the PCR-free method is also in line with the application of multigene-based environmental assessment, which can improve the accuracy of biodiversity detection (David *et al.* 2012). However, an obvious challenge in bypassing PCR is that the procedure is dependent on the original proportion of mitochondrial DNA in total genomic DNA, which is typically at the scale of 0.5% in insects. Although much more genomic DNA may possess taxonomic resolution (e.g. many nuclear genes), current molecular identification

Correspondence: Xin Zhou, Fax: +86 0755 2235 4236; E-mail: xinzhou@genomics.cn

systems for animals are primarily based on mitochondrial markers. Conventional mitochondrial enrichment approaches (e.g. differential centrifugation) were ineffective in increasing mitochondrial ratio in pooled arthropod samples [e.g. <0.5% useful mitochondrial reads reported in (Zhou *et al.* 2013)], therefore inevitably resulting in a huge excess of data and elevation of operational costs. In addition, the massive volume of sequence data is highly demanding on computing resources. Thus, the implementation of a PCR-free shotgun approach in routine biodiversity analysis will benefit from a significant improvement in sequencing efficiency, which can be achieved by increasing the proportion of mitochondrial DNA in the DNA soup.

Target-region enrichment, as it is routinely adopted in, for example, exon capture of human DNA (Bamshad *et al.* 2011), ancient human DNA baiting (Burbano *et al.* 2010), enrichment of pathogen DNA (Devault *et al.* 2014) or highly degraded DNA from museum vouchers (Guschanski *et al.* 2013), offers a cost-effective method for enriching mitochondrial genome. A few studies reported success in cross-taxon sequence hybridization: for instance, a human exon microarray was able to capture approximately 95% of genomes from nonhuman primates (Vallender 2011) and mitochondrial genomes of 13 colugo species were successfully captured from museum specimens (Mason *et al.* 2011). Furthermore, studies also showed possibilities in capturing conserved genes among highly divergent species (Lemmon *et al.* 2012; Li *et al.* 2013), albeit with relatively low capture efficiency. However, all these studies aimed at capturing genes from one species at a time or from mixed DNA libraries with known tags (Hancock-Hanser *et al.* 2013), rather than from mixed mass or environmental samples containing multiple species. Expected challenges in developing a probe-based mitochondrial capture array targeting a wide range of taxa include the lack of a comprehensive reference library of mitochondrial sequences covering every taxonomic lineage that may be encountered in various natural habitats and the lack of optimization in hybridization conditions. It is not clear whether probes designed with reference genomes could successfully capture many species in mass samples or environmental DNAs that are related to the references at various levels of phylogenetic distance.

In the present study, we designed a mitochondrial capture microarray by capitalizing on mitochondrial genes of 379 species produced by the 1000 Insect Transcriptome Evolution project [1KITE, www.1kite.org, (Misof *et al.* 2014)]. An assemblage of DNA containing 49 species identified in a previous study (Tang *et al.* 2014) was utilized to evaluate the mitochondrial enrichment competency of the newly designed array (Table 1). We explored whether known input species could be captured by a set of mitochondrial probes designed based on nontarget taxa. We also examined whether target taxa of varied phylogenetic distances to probes would affect capture efficiency in terms of length coverage and relative abundance. Our study illustrates the potential of applying a mitochondrial capture microarray pipeline in bulk environmental samples before PCR-free shotgun sequencing.

Material and methods

The design of capture microarray

A total of 2553 mitochondrial protein-coding scaffolds with an average length of 1902 bp were extracted by gene annotation (Zhou *et al.* 2013) from 379 1KITE transcriptomes including all extant insect orders as well as outgroups, that is Maxillopoda, Malacostraca, Myriapoda, Remipedia and Crustacea (Fig. 1). The 1KITE data set adds more than 100 new families (mainly in Odonata, Plecoptera, Dermaptera, Mantodea and Hymenoptera [Table S2]) to the insect groups that are already represented by published mitogenomes (Cameron 2014). These new families are evenly distributed across the phylogenetic tree of insects. All mitochondrial scaffolds were used for the following probe design and synthesis, encompassing four major steps (Fig. S1):

Probe length selection. First, three sets of K-mer sequences were inferred from mitochondrial scaffolds with K-length of 65, 73 and 93 bp, respectively. A K-mer set was obtained by breaking all mitochondrial scaffolds into k-mers in a single step. Then, theoretical melting temperatures were calculated for each K-mer set (SantaLucia 1998). We finally selected 73-mers for the following analysis based on the level of consistency in melting temperatures and the mismatch tolerance.

Table 1 The taxonomic composition of the mock sample

Phylum	Class	Order	# of species
Arthropoda	Arachnida	Araneae	1
		Opiliones	1
	Branchiopoda	Cladocera	1
	Insecta	Blattodea	1
		Coleoptera	4
		Diptera	5
		Embioptera	1
		Ephemeroptera	7
		Hemiptera	5
		Hymenoptera	4
		Lepidoptera	13
		Odonata	2
		Orthoptera	2
Chordata	Actinopterygii	Cypriniformes	1
Echinodermata	Asteroidea	Forcipulatida	1

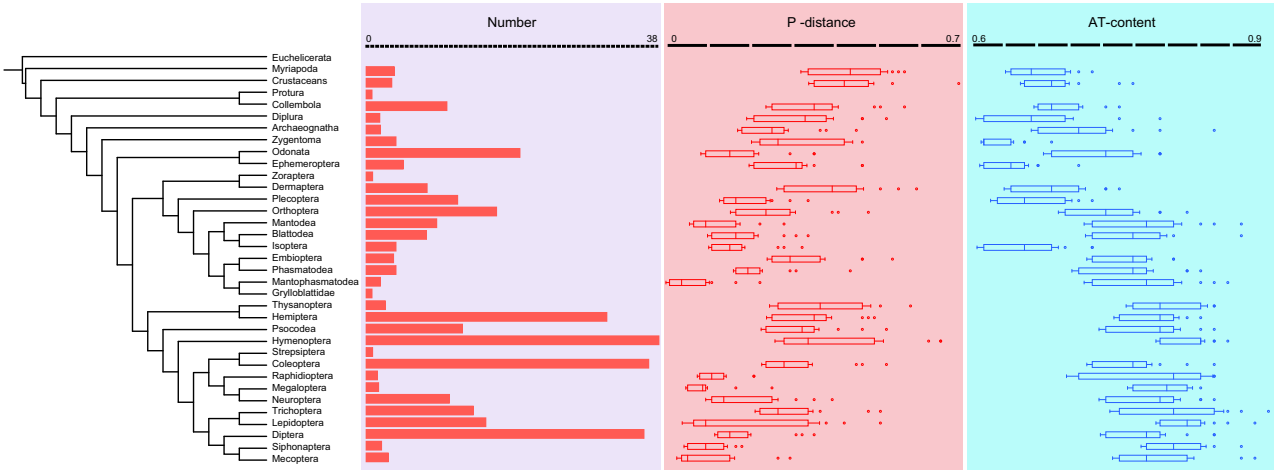

Fig. 1 Summary of mitochondrial protein-coding genes in 379 references. The number represents the quantity of mitochondrial genomes in each group, which is calculated by the number of protein-coding genes divided by 13; the p-distance represents the average value of each protein-coding gene within groups; and the AT content also represents the average value of each protein-coding gene per group.

Probe score evaluation. A comprehensiveness score for each 73-mer sequence was calculated based on four parameters, including (i) level of uniqueness; (ii) level of frequency; (iii) theoretical melting temperature; and (iv) GC content. Parameters used in evaluating each 73-mer sequence and the detailed algorithm in calculating the comprehensiveness score value are provided in Appendix S1.

Probe design. Regions were selected from mitochondrial scaffolds with a sliding window of 45 bp at a 20-bp interval. Then, 73-mers belonging to each region were sorted by their comprehensiveness scores, and the one with highest score was chosen as the probe for this region (Appendix S1). Additionally, the quantity of probes with GC content >60% was doubled in the final probe set to improve the capture efficiency in GC-rich regions, which tend to form self-hairpin structures, leading to low capture efficiency (Lemoine *et al.* 2009).

Probe synthesis. The designed probes were imported into a CustomArray B3™ Synthesizer (CustomArray, Washington, DC, USA) to produce DNA oligonucleotides following the manufacturer's protocols. Oligonucleotides were washed out and collected by concentrated ammonia. Finally, a total of 187 674 RNA probes were generated using PCR and reverse transcription.

Mock DNA soup, hybridization and sequencing

The pooled DNA sample was derived from our recent study (Tang *et al.* 2014) containing 49 animal species from 47 genera and 42 families (all but two from Arthropoda), with most taxa representing a single family (Table 1). Genomic DNA from each specimen was individually extracted following Ivanova *et al.* (2006). Aliquots of DNA (each containing 100 ng of gDNA) from 49 species (Table 1) were pooled for library construction with an insert size of 200 bp. DNA hybridization to the microarray was conducted as following. Briefly, DNA was fragmented using an ultrasonoscope (Covaris S2, MA, USA), and the fragments then underwent end-repair, a single 'A' base addition, adapter ligation and size selection on an agarose gel (200 ± 20 bp). Next, the product was PCR-amplified by 4 cycles to confirm the adapter ligation. DNA with adapters were mixed together with probes to hybridize for 72 h. Probes applied were in the same quantity as that of total DNA, which was thousands of times more than mitochondrial sequences, thus offering excessive probes for effective capture. Then, streptavidin-coated magnetic beads were added to the mixture to match with the biotin on probes via noncovalent interactions. Finally, uncaptured DNA fragments were rinsed out, and captured DNA was sent for sequencing on an Illumina HiSeq-2000 at BGI-Shenzhen with a strategy of 100 PE (paired-end) sequencing producing 2 Gbp of raw data after adapter removal.

Test of capture efficiency

P-distance and AT content evaluation for probes. In principle, a probe-based capturing protocol performs best for taxa closely related to reference taxa from which probes have been designed. To understand the effect of phylogenetic relatedness on capture efficiency, we calculated

p-distances of homologous sequences between target taxa and reference taxa used in probe design. We used Clustalw (Larkin *et al.* 2007) for full-length alignments for homologous genes with a penalty value of 50 for gap opening and extension. Then, p-distances and AT contents were calculated using in-house Perl scripts (Appendix S2). For mitochondrial genes used in probe design, average p-distance and AT content of each protein-coding genes (PCG) were calculated for each order or group (Fig. 1) to illustrate the distribution of probes across arthropod lineages and their corresponding sequence characteristics.

Reference mitogenomes of the mock sample. The mitochondrial genomes of our 49 species are available from Tang *et al.* (2014). Briefly, DNA was extracted individually from the 49 species, pooled, fragmented and directly sequenced on a Hiseq-2000 without PCR amplification or enrichment, generating a total of 35 Gb of data. TGICL (Pertea *et al.* 2003) was used to combine assembled scaffolds generated by SOAPdenovo (Luo *et al.* 2012), SOAPdenovo-Trans (Xie *et al.* 2014) and IDBA_UD (Peng *et al.* 2012). Finally, concatenated mitochondrial scaffolds were assigned to input species based on sequence homology to public databases (e.g. BOLD, NCBI) and local reference Sanger sequences. We identified all 13 protein-coding genes for all 49 species except for *Aposthionia borneensis* (missing CYTB) and *Opiliones* (missing ND4, ND4L, ND6 and CYTB) (Tang *et al.* 2014). The most closely related homologues in the probe set to each PCG of the target taxa were identified, and p-distance was calculated using following steps: (i) each PCG was BLASTed against all probes to find homologues of high similarity; (ii) the top 5 probe sequences were selected for each PCG and subject to full-length alignment using Clustalw (Larkin *et al.* 2007) with a penalty value of 50 for gap opening and extension; and (iii) the probe sequence with the smallest p-distance was identified as the closest homologue to the corresponding PCG, and this p-distance was used in the following analysis.

Capture efficiency and relative abundance. BWA (Li & Durbin 2009) was employed to map raw reads to reference mitogenomes by allowing for ≤2 mismatches. The capture efficiency was evaluated by the coverage percentage of each of the 49 reference mitogenomes. Only regions mapped by >3 reads were considered as valid coverage, considering that fragments of mitochondrial genomes can be randomly sequenced even without capture enrichment. The average abundance, measured by average sequence depth, was calculated as the anchored read number weighted by corresponding regional length. To understand the influence of p-distance and AT content on capturing, reference genomes were divided into two categories: regions that can be covered by captured reads and regions that cannot. Then, p-distance and AT content of these two categories were conducted for variance analysis to find out whether significant discrepancies can be detected between them. Finally, correlation analysis was carried out between coverage rate of each species and their initial abundance and p-distance.

The influence of reference comprehensiveness on species detection. Different from the PCR-based, single-target-marker metabarcoding approach, PCR-free mitogenomics can take into consideration the coverage ratio of reference mitogenomes for determining the confidence of species presence/absence. In principle, taxa absent from the bulk sample would receive low coverage for their reference mitogenomes. An additional 3768 mitochondrial genomes of Arthropods, Echinodermata and Chordata obtained from GenBank were integrated with our 49 genomes serving as reference genomes to test whether nontarget references will gain as high coverage as the 49 species present in the mock sample, especially for references whose closely related taxa are present in the mock sample.

The influence of capture on abundance. Zhou *et al.* (2013) showed that the number of mitochondrial nucleotides and biomass (predicted by body length) were correlated in pooled arthropod samples using a PCR-free based shotgun sequencing approach. This observation implies that relative abundance of species may be estimated from mixed NGS samples (Gómez-Rodríguez *et al.* 2015; Tang *et al.* 2015). The average sequence depths of the 49 input taxa before probe capture were obtained from Tang *et al.* (2014, Table 1) serving as initial abundances, while those for after capture were calculated in this study. We calculated the *D*-values (discrepancy of abundance) of each species between initial abundance and that of after capture, and correlation analysis was conducted between these two data sets. In addition, species were divided into groups according to their p-distances. *D*-value of each group was then evaluated to identify the potential influence of p-distance on abundance alteration after capture.

Results

Summary of probes

We mapped the assembled mitochondrial fragments of 379 1KITE species onto the published insect backbone tree (Misof *et al.* 2014). The number of mitogenomes, average p-distance (dissimilarity within group) and average AT content were recorded for each insect order (Fig. 1, Table S3). The number of mitogenomes averaged 9.64 for each insect order with the maximum in

Hymenoptera (38) and a minimum of 1 in Protura, Strepsiptera, Zoraptera and Grylloblattodea. Myriapods and crustacean were the two most divergent groups in our mitogenome reference data set, and Mantophasmatodea was the least diverged group. The AT content was highest in Hymenoptera and lowest in Zygentoma.

Capture efficiency

Proportions of reads that can be mapped to the 49 reference mitogenomes were compared between pooled DNA samples of before- and after capture, revealing a large increase from 0.47% to 42.52%. The high capture efficiency was also demonstrated by the success in hybridization of some highly variable mitochondrial sequences from the pooled DNA, for example regions of the reference mitogenomes that were only 70% identical to the probes were still captured (Fig. S2). As illustrated in Fig. 2, coverage rate (length coverage) for nearly all insects was >80%, except in three hymenopterans (*Polyrhachis dives*, Vespidae and Ichneumonidae at 72%, 69% and 52%, respectively), one hemipteran (*Laternaria candelaria*, 71%) and one embiopteran (*Aposthonia borneensis*, 43%). Although the

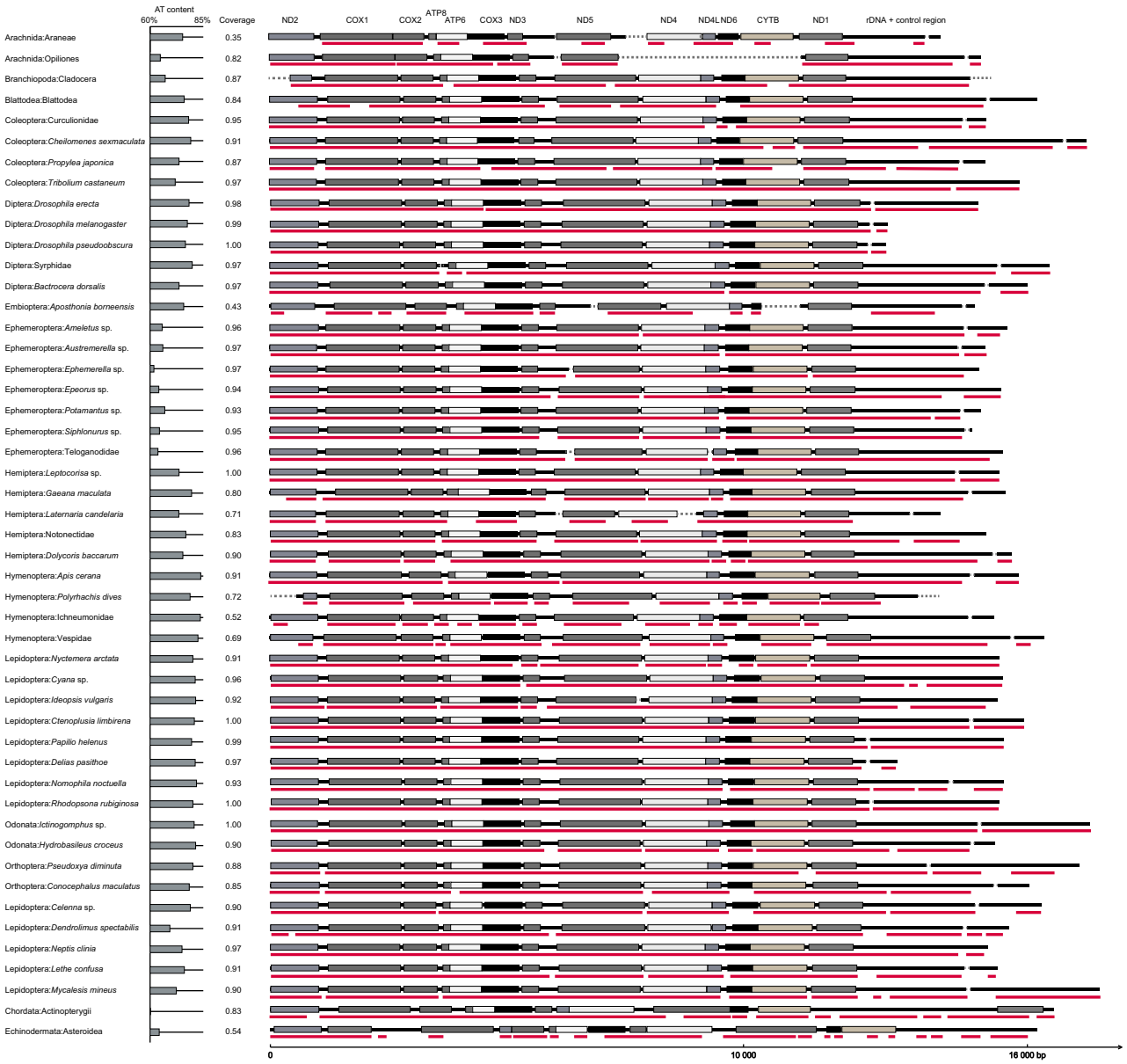

Fig. 2 The mitochondrial genome coverage of each species after gene capture. Bars of different greyscales represent the mitochondrial genomes of the references, and red lines represent the corresponding regions that were covered by reads after gene capture.

probe set was designed primarily based on insect transcriptomes, more than half of the mitochondrial genome sequences of starfish and spiders were recovered, and 83% of the zebrafish mitogenome was recovered.

Mitochondrial regions that failed to be captured were examined for both dissimilarity to probes and AT content and were compared to those of the successfully captured regions. ANOVA showed that the failed regions had marginally significantly higher AT content ($P = 0.068$, d.f. = 1) and significantly higher p-distances to probes ($P < 0.01$, d.f. = 1). Correlation analyses were also conducted for coverage rate vs. initial abundance and coverage rate vs. p-distance for each species. Both factors showed significant correlations ($P < 0.01$) to coverage rate, with $r = 0.52$ and -0.55, respectively.

More than half of the captured sequences were not derived from any of the mock-sample taxa. Raw reads of these nontarget sequences (those that could not be aligned to references) were assembled using SOAPdenovo-Trans (Xie *et al*. 2014). Assemblies with length >1000 bp were calculated for p-distances against the probe set. The results showed that only a few nontarget sequences (73 of 1161, 6.29%) found hits in the probe set using blast with the e parameter of 1e-5, indicating high dissimilarity. Therefore, probes are more likely to hybridize with DNA fragments of high similarity, but they will also capture DNA fragments randomly regardless of sequence similarity when more similar targets are absent. Alternatively, part of these nontarget raw reads could be fragments that were not hybridized but not washed away after capture.

Species detection

Species detection via whole mitogenome shotgun sequencing was made based on coverage (length proportion) of reference mitogenomes. When species present in the bulk sample (target taxa) were represented in the reference, the corresponding reference mitogenome typically received a high coverage, from which solid conclusion on species presence was drawn. For instance, all 49 taxa from the mock sample exhibited coverages of >50%, and 46 of them exhibited >75% coverage (Table S4). Among these 49 taxa, species with remote distance to probes tend to yield relatively lower coverage, for example spiders and starfish in our mock sample showed approximately 50% reference coverage (Fig. 2, Table S4). In cases where a series of closely related taxa were included in the reference, those present in the mock sample gained significantly higher coverage compared to those absent. For example, *Drosophila erecta*, *Drosophila melanogaster* and *D. pseudoobscura* were in the mock sample, and each had coverage of >98%, but *D. mauritiana*, which was absent in the mock sample, had only 16.2% coverage (Table S4).

Relative abundance

The average sequence depth of each species showed an increase of an average of 5.8-fold (standard deviation of 2.2) after capture, which is consistent with the volume discrepancy between the two data sets: 42.52% mitochondrial reads in 2 Gbp raw data after capture vs. 0.47% mitochondrial reads of 35 Gbp raw data without capture ($42.52/0.47 \times 2/35 = 5.17$). As illustrated in Fig. 3A, no significant changes in overall relative abundance were observed after capture. In fact, a strong positive correlation was found in read abundance between before and after applying the capture technique and sequencing ($r^2 = 0.81$, Fig. 3B). *D*-values of each species were calculated as: *D*-value = a/5.8 – b (a = abundance after capture, b = initial abundance) and then tested for correlation with p-distances, leading to nonsignificant results. Samples were then divided into three groups: groups of p-distance <0.2, 0.2–0.35 and >0.35. The numbers of samples with *D*-value >0 and *D*-value <0 were counted for each category. Fisher exact count number test demonstrated these three groups were significantly different ($P < 0.01$, Fig. 4).

In summary, these observations indicated that, for the given experimental set-up, the probe capture protocol significantly increased mitochondrial reads for pooled taxa by approximately 100X, and all tested known taxa were captured by our pipeline. Species detection was affected by both the phylogenetic relatedness of target taxa to probe set and the level of comprehensiveness of the reference genome database. The relative abundance of each input species was subject to variants in phylogenetic distances of target taxa to probe set and initial abundance. However, the overall abundance of input taxa was largely maintained after capture.

Discussion

In the past few years, scientists have dedicated tremendous efforts to construct references for carefully selected DNA markers. For instance, the International Barcode of Life (iBOL, www.ibol.org) project has come a long way in the accumulation of DNA barcode sequences, including CO1 as the animal barcode marker, reaching a number of 4 681 401, covering approximately 0.24 million species (BOLD, accessed in June 2015). Several approaches have been proposed to improve the overall efficiency in metabarcoding

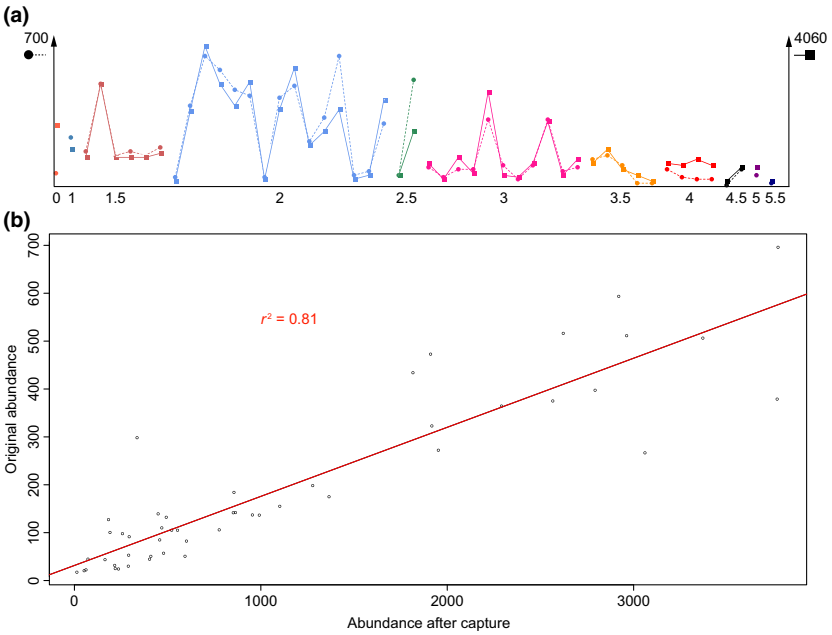

Fig. 3 Correlation between initial abundance and the abundance after capture. A: Squares connected by full lines represent initial abundance obtained from Tang *et al.* (2014), corresponding to the value on the left vertical axis. Circles connected by dashed lines represent abundance after gene capture, corresponding to the value on the right vertical axis. The horizontal axis represents the p-distance to their most similar probes of each species. B: correlation test of the abundance of each species.

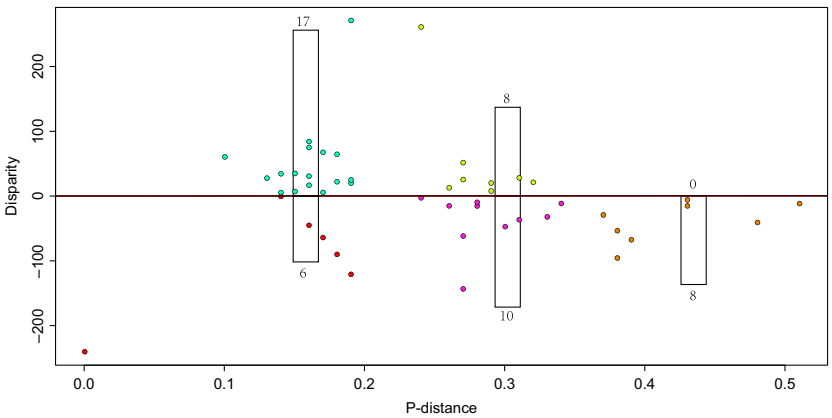

Fig. 4 Correlation between *D*-value and p-distance. The horizontal axis represents the p-distance to their most similar probes of each species. The vertical axis represents *D*-values of each species. The height of each bar corresponds to the number *D*-value >0 (upside) or *D*-value <0 (downside). Circles with identical colours belong to the same groups.

pipelines, including the use of a fragment of standard barcodes [e.g. 'mini-barcode' (Hajibabaei *et al.* 2006)], new markers designed specifically for environmental analysis (Riaz *et al.* 2011) and PCR-free methods (Zhou *et al.* 2013). In addition to the standard CO1 barcode, other mitochondrial genes or whole mitogenomes may also serve as effective references in animal species delineation. Studies aiming to construct mitochondrial genome references in high-throughput manners have also been conducted using NGS platforms (Timmermans *et al.* 2010; Groenenberg *et al.* 2012; Rubinstein *et al.* 2013; Gillett *et al.* 2014; Tang *et al.* 2014; Williams *et al.* 2014). Mitochondrial capture will be complementary to existing reference-based PCR-free methods, to the extent that it can enrich mitogenomes without introducing significant taxonomic biases for mixed DNA samples.

Mitochondrial DNA of pooled species can be captured efficiently using cross-taxa probes

Our method represents the first effort to simultaneously capture different mitochondrial DNA from mass samples to recover taxonomic composition. Although previous work has demonstrated some success in cross-taxa sequence capture (Lemmon *et al.* 2012; Li *et al.* 2013), it is crucial to understand whether and how the presence of multiple divergent species affect capture success. Our results suggest that probes based on references from a wide range of insect lineages are efficient in capturing mitochondrial DNA of pooled species, even though none of the species in the mock sample was used for probe design. It is reasonable to assume that a large proportion of the 'wanted' taxa that need to be enriched by the mitochondrial capture procedure will not be present in the

array design, at least not on the species level. In fact, probe sequences will never be comprehensive and complete despite the fact that full mitochondrial genomes are being produced at a fast pace. Therefore, our observations are encouraging because a probe set that contains representatives from major lineages of target organisms can capture most, if not all, species from pooled environmental samples.

The capture pipeline also picked up nontarget DNA, for example 57.48% of hybridized sequences in this study, most of which are presumably nuclear DNA. As suggested in the results, both p-distance and initial abundance significantly affected the capture rate (coverage of reference genome) of each input species. When input species were categorized into two groups based on their p-distances, binary linear regression analysis demonstrated that the impact of initial abundance increased in samples with high p-distance (Appendix S3). This discovery explains why regions of the target mitogenomes at low similarity to probes and nuclear genes, which are present in high amounts and proportion, could still be partially captured by the pipeline.

Species richness and relative abundance can be achieved for biodiversity assessment

The taxonomic composition in the mock sample can be recovered by examining coverage rates of reference mitogenomes. Taxa can be easily detected from bulk sample at high coverage when they are present in the reference. Reference taxa absent from bulk sample sometimes may also obtain low coverage when closely related species are present in the sample. These can be ruled out by the significant discrepancies in coverage rates compared to true positives. Obviously, species absent from the reference cannot be detected using our method. Therefore, the current pipeline is more suitable for locations where well-established reference is available, or studies focusing on certain important indicative taxa. As for other reference-based approaches, we recommend construction of comprehensive reference libraries of mitogenomes for studied habitats or organisms, which will obviously benefit from a series of recent technological developments (Timmermans *et al.* 2010; Groenenberg *et al.* 2012; Rubinstein *et al.* 2013; Gillett *et al.* 2014; Williams *et al.* 2014), including those from our own work (Tang *et al.* 2014, 2015).

Compared to the success in recovering species richness, the capture procedure, as an extra step to the PCR-free approach, is more likely to produce biases to relative abundance (represented by sequence depth) of each target species. Indeed, our analysis showed that the capture procedure did seem to pose different impacts on taxa possessing varied phylogenetic distances to probes. The discrepancy of abundances (*D*-values) of species with smaller p-distances tended to be smaller than those with larger p-distances, implying taxa that are phylogenetically more distant to those used in probe design are inclined to lose more mitochondrial DNA during the capture (Fig. 4). However, the overall capture result showed that relative abundances of before and after capture for input species had a strong correlation ($r^2 = 0.81$, Fig. 3B). Therefore, we suspect that the capture bias caused by varied phylogenetic relatedness of target taxa to probes may not significantly alter ecological inferences, considering the prospect that the capture array can be expanded to cover more representative lineages, or alternatively, customized to any given fauna. Apparently, systematic test on the ecological applications is much needed to optimize the capture pipeline.

Mitochondrial capture as an economical pretreatment for PCR-free based biodiversity analysis

Operational cost is a game changer in consideration of any new biodiversity applications. The major expenditure of liquid capture array is on probe synthesis, which is approximately 6000 USD per 96K chip in the present study. The capacity of each synthesis is sufficient for >2000 captures, resulting in approximately 3 USD per sample. Although the capture procedure will not remove the cost of library construction, it will greatly reduce the demand of sequencing volume for a PCR-free based shotgun analysis. For instance, for bulk arthropod samples with similar richness (approximately 40–50 spp.), the sequencing volume of this study is <15% of that in Zhou *et al.* (2013).

Summary and Perspective

Current pilot capture work combined with our former studies with regards to mitochondrial reference construction (Tang *et al.* 2014) and PCR-free biomonitoring (Zhou *et al.* 2013; Tang *et al.* 2015) will open up new and exciting opportunities for biodiversity assessment. With the rapid accumulation of mitogenomes and robust reference of specific environments or groups, such methods would be applied to several important fields in the foreseeable future, such as quarantine inspection, aquatic and agriculture ecosystems scrutiny.

Acknowledgements

We are thankful to Douglas Yu for his valuable advices. This work is supported by the National High Technology Research and Development Program of China – 863 Program (2012021601), the National Key Technology R&D Program

(2012BAK11B06) and the Science and Technology Innovation of CAS, iFlora Cross and Cooperation Team (31129001). We would like to thank members of the 1KITE consortium (www.1kite.org) for allowing us to filter mitogenomes in transcriptome data.

References

Baird DJ, Hajibabaei M (2012) Biomonitoring 2.0: a new paradigm in ecosystem assessment made possible by next-generation DNA sequencing. *Molecular Ecology*, **21**, 2039–2044.

Bamshad MJ, Ng SB, Bigham AW *et al.* (2011) Exome sequencing as a tool for Mendelian disease gene discovery. *Nature Reviews Genetics*, **12**, 745–755.

Bienert F, Dedanieli S, Miquel C *et al.* (2012) Tracking earthworm communities from soil DNA. *Molecular Ecology*, **21**, 2017–2030.

Board MA (2005) *Millennium Ecosystem Assessment*. New Island, Washington, DC.

Burbano HA, Hodges E, Green RE *et al.* (2010) Targeted investigation of the Neandertal genome by array-based sequence capture. *Science*, **328**, 723–725.

Cameron SL (2014) Insect mitochondrial genomics: implications for evolution and phylogeny. *Annual Review of Entomology*, **59**, 95–117.

Coissac E, Riaz T, Puillandre N (2012) Bioinformatic challenges for DNA metabarcoding of plants and animals. *Molecular Ecology*, **21**, 1834–1847.

David O, Larédo C, Leblois R, Schaeffer B, Vergne N (2012) Coalescent-based DNA barcoding: multilocus analysis and robustness. *Journal of Computational Biology*, **19**, 271–278.

Devault AM, McLoughlin K, Jaing C *et al.* (2014) Ancient pathogen DNA in archaeological samples detected with a Microbial Detection Array. *Scientific Reports*, **4**, e4245.

Gillett CP, Crampton-Platt A, Timmermans MJ *et al.* (2014) Bulk de novo mitogenome assembly from pooled total DNA elucidates the phylogeny of weevils (Coleoptera: Curculionoidea). *Molecular Biology and Evolution*, **31**, 2223–2237.

Gómez-Rodríguez C, Crampton-Platt A, Timmermans MJ, Baselga A, Vogler AP (2015) Validating the power of mitochondrial metagenomics for community ecology and phylogenetics of complex assemblages. *Methods in Ecology and Evolution*, **6**, 883–894.

Groenenberg DS, Pirovano W, Gittenberger E, Schilthuizen M (2012) The complete mitogenome of Cylindrus obtusus (Helicidae, Ariantinae) using Illumina next generation sequencing. *BMC Genomics*, **13**, 114.

Guschanski K, Krause J, Sawyer S *et al.* (2013) Next-generation museomics disentangles one of the largest primate radiations. *Systematic Biology*, **62**, 539–554.

Hajibabaei M, Smith M, Janzen DH *et al.* (2006) A minimalist barcode can identify a specimen whose DNA is degraded. *Molecular Ecology Notes*, **6**, 959–964.

Hajibabaei M, Shokralla S, Zhou X, Singer GAC, Baird DJ (2011) Environmental barcoding: a next-generation sequencing approach for biomonitoring applications using river benthos. *PLoS One*, **6**, e17497.

Hancock-Hanser BL, Frey A, Leslie MS *et al.* (2013) Targeted multiplex next-generation sequencing: advances in techniques of mitochondrial and nuclear DNA sequencing for population genomics. *Molecular Ecology Resources*, **13**, 254–268.

Ivanova NV, Dewaard JR, Hebert PD (2006) An inexpensive, automation-friendly protocol for recovering high-quality DNA. *Molecular Ecology Notes*, **6**, 998–1002.

Ji Y, Ashton L, Pedley SM *et al.* (2013) Reliable, verifiable and efficient monitoring of biodiversity via metabarcoding. *Ecology Letters*, **16**, 1245–1257.

Keesing F, Belden LK, Daszak P *et al.* (2010) Impacts of biodiversity on the emergence and transmission of infectious diseases. *Nature*, **468**, 647–652.

Larkin MA, Blackshields G, Brown N *et al.* (2007) Clustal W and Clustal X version 2.0. *Bioinformatics*, **23**, 2947–2948.

Lemmon AR, Emme SA, Lemmon EM (2012) Anchored hybrid enrichment for massively high-throughput phylogenomics. *Systematic Biology*, **61**, 727–744.

Lemoine S, Combes F, Le Crom S (2009) An evaluation of custom microarray applications: the oligonucleotide design challenge. *Nucleic Acids Research*, **37**, 1726–1739.

Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, **25**, 1754–1760.

Li C, Hofreiter M, Straube N, Corrigan S, Naylor GJ (2013) Capturing protein-coding genes across highly divergent species. *BioTechniques*, **54**, 321–326.

Liu S, Li Y, Lu J *et al.* (2013) SOAPBarcode: revealing arthropod biodiversity through assembly of Illumina shotgun sequences of PCR amplicons. *Methods in Ecology and Evolution*, **4**, 1142–1150.

Luo R, Liu B, Xie Y *et al.* (2012) SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *GigaScience*, **1**, 18.

Mason VC, Li G, Helgen KM, Murphy WJ (2011) Efficient cross-species capture hybridization and next-generation sequencing of mitochondrial genomes from noninvasively sampled museum specimens. *Genome Research*, **21**, 1695–1704.

Misof B, Liu S, Meusemann K *et al.* (2014) Phylogenomics resolves the timing and pattern of insect evolution. *Science*, **346**, 763–767.

Naidoo R, Weaver LC, Stuart-Hill G, Tagg J (2011) Effect of biodiversity on economic benefits from communal lands in Namibia. *Journal of Applied Ecology*, **48**, 310–316.

Peng Y, Leung HC, Yiu S-M, Chin FY (2012) IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics*, **28**, 1420–1428.

Pertea G, Huang X, Liang F *et al.* (2003) TIGR Gene Indices clustering tools (TGICL): a software system for fast clustering of large EST datasets. *Bioinformatics*, **19**, 651–652.

Porazinska DL, Giblinavis RM, Faller L *et al.* (2009) Evaluating high throughput sequencing as a method for metagenomic analysis of nematode diversity. *Molecular Ecology Resources*, **9**, 1439–1450.

Porazinska DL, Sung W, Giblin-Davis RM, Thomas WK (2010) Reproducibility of read numbers in high-throughput sequencing analysis of nematode community composition and structure. *Molecular Ecology Resources*, **10**, 666–676.

Riaz T, Shehzad W, Viari A *et al.* (2011) ecoPrimers: inference of new DNA barcode markers from whole genome sequence analysis. *Nucleic Acids Research*, **39**, e145.

Rubinstein ND, Feldstein T, Shenkar N *et al.* (2013) Deep sequencing of mixed total DNA without barcodes allows efficient assembly of highly plastic ascidian mitochondrial genomes. *Genome Biology and Evolution*, **5**, 1185–1199.

SantaLucia J (1998) A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics. *Proceedings of the National Academy of Sciences*, **95**, 1460–1465.

Taberlet P, Coissac E, Pompanon F, Brochmann C, Willerslev E (2012) Towards next-generation biodiversity assessment using DNA metabarcoding. *Molecular Ecology*, **21**, 2045–2050.

Tang M, Tan M, Meng G *et al.* (2014) Multiplex sequencing of pooled mitochondrial genomes-a crucial step toward biodiversity analysis using mito-metagenomics. *Nucleic Acids Research*,**42**, e166.

Tang M, Hardman C, Ji YQ *et al.* (2015) High-throughput monitoring of wild bee diversity and abundance via mitogenomics. *Methods in Ecology and Evolution*, **6**, 1034–1043

Timmermans MJ, Dodsworth S, Culverwell C *et al.* (2010) Why barcode? High-throughput multiplex sequencing of mitochondrial genomes for molecular systematics. *Nucleic Acids Research*, **38**, e197.

Vallender EJ (2011) Expanding whole exome resequencing into non-human primates. *Genome Biology*, **12**, R87.

Williams S, Foster P, Littlewood D (2014) The complete mitochondrial genome of a turbinid vetigastropod from MiSeq Illumina sequencing of genomic DNA and steps towards a resolved gastropod phylogeny. *Gene*, **533**, 38–47.

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

Yu DW, Ji YQ, Emerson BC *et al*. (2012) Biodiversity Soup: metabarcoding of arthropods for rapid biodiversity assessment and biomonitoring. *Methods in Ecology and Evolution*, **3**, 613–623.

Zhou X, Li Y, Liu S *et al*. (2013) Ultra-deep sequencing enables high-fidelity recovery of biodiversity for bulk arthropod samples without PCR amplification. *GigaScience*, **2**, 4.

S.L. and X.Z. designed the study; X.W., S.L., J.H. and X.Z. contributed to the project coordination; S.L. led the analyses, L.X., Z.L., Z.H. and M.T. conducted the design and synthesis of probes. X.S. and M.I.T. contributed to the bench work; O.N. helped to provide part of transcriptome samples; S.L. and X.W. wrote the first draft, and X.Z., B.M., O.N. and K.K. contributed to revisions.

Data Accessibility

Probes and ClustalW alignments: Dryad entry doi:10.5061/dryad.19nf6.

Raw reads: NCBI SRA: SRX1202771.

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Fig. S1 Flow diagram of probe design.

Fig. S2 The regional mappable read number (upper-panel) and their corresponding p-distance (lower panel) for each of the 49 species from our mockup sample.

Table S1 The taxonomic information of the mock sample.

Table S2 Taxonomic comparison between public data and 1KITE.

Table S3 Taxonomic and mitochondrial gene information of 1KITE.

Table S4 Coverage represented the proportion of the reference genomes that has been mapped by raw reads;species made up mockup samples has been marked by green blow.

Appedix S1 Algorithm of comprehensiveness score calculation in probe design.

Appedix S2 In-house Perl scripts.

Appedix S3 Binary linage analysis of the influence of P-distance and abundance on coverage rate.