bioRxiv preprint doi: https://doi.org/10.1101/080689; this version posted October 14, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

An improved *de novo* pipeline for enrichment of high diversity mitochondrial genomes from Amphibia to high-throughput sequencing

Xing Chen[1†], Gang Ni[1,2†], Kai He[1†], Zhao-Li Ding[3,4†], Gui-Mei Li[3,4], Adeniyi C. Adeola[1,5,6], Robert W. Murphy[1,7], Wen-Zhi Wang[1,6,8*], Ya-Ping Zhang[1,2,6,9*]

[1] State Key Laboratory of Genetic Resources and Evolution, Kunming Institute of Zoology, Chinese Academy of Sciences, No. 32 Jiaochang Donglu, Kunming, Yunnan 650223, China, [2] Yunnan Laboratory of Molecular Biology of Domestic Animals, Kunming Institute of Zoology, Chinese Academy of Sciences, No. 32 Jiaochang Donglu, Kunming, Yunnan 650223, China, [3] Kunming Biological Diversity Regional Centre of Large Apparatus and Equipments, Kunming Institute of Zoology, Chinese Academy of Sciences, Kunming, Yunnan 650223, China, [4] Public Technology Service Centre, Kunming Institute of Zoology, Chinese Academy of Sciences, No. 32 Jiaochang Donglu, Kunming, Yunnan 650223, China, [5] China-Africa Centre for Research and Education & Yunnan Laboratory of Molecular Biology of Domestic Animals, Kunming Institute of Zoology, Chinese Academy of Sciences, No. 32 Jiaochang Donglu, Kunming, Yunnan 650223, China, [6] Animal Branch of the Germplasm Bank of Wild Species, Kunming Institute of Zoology, Chinese Academy of Sciences, No. 32 Jiaochang Donglu, Kunming, Yunnan 650223, China, [7] Centre for Biodiversity and Conservation Biology, Royal Ontario Museum, 100 Queen's Park Toronto, Ontario M5S 2C6, Canada, [8] Forensic Science Service of Yunnan Endangered Species Scientific Commission, No. 32 Jiaochang Donglu, Kunming, Yunnan 650223, China, [9] Laboratory for Conservation and Utilization of Bio-resource and Key Laboratory for Microbial Resources of the Ministry of Education, Yunnan University, Kunming, Yunnan 650091, China

* To whom correspondence should be addressed. Tel: +86-871-68125502; Fax: +86-871-68125513; E-mail: wangwz@mail.kiz.ac.cn, Tel: +86-871- 68526504; Fax: +86-871- 68526504; E-mail: zhangyp@mail.kiz.ac.cn

†These authors contributed equally to this work.

ABSTRACT

We present an improved de novo pipeline by combining long-range polymerase chain reaction (LR-PCR) and capture hybridization for enriching mitochondrial DNA to high-throughput sequencing. We test a new set of primers and hybridizing long-range library (LR-HY) with 112 mitochondrial genomes (MtG) representing three orders, 12 families, 54 genera, and 106 species of Amphibia. The primers are used for obtaining wide taxonomic MtG amplicons to sequence directly and/or make probes for closely related species. LR-HY is compared to standard hybridization. The primers successfully amplify 82 MtGs from all three order, all families, 92.6% (50/54) of the genera, and 74.5% (79/106) of the species, despite some

bioRxiv preprint doi: https://doi.org/10.1101/080689; this version posted October 14, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

DNA degradation and gene rearrangement. We observe a significantly negative correlation between sequence depth and gene variation. The pattern of highly variable regions is separately distributed in different regions within the length of < 4 kb in the 33-pooled sample. We demonstrate that using 2 kb libraries generate deeper sequence coverage in the highly variable loci than using 400 bp libraries. In total, the pipeline successfully recovers 83 complete and 14 almost complete MtGs from 53 of 54 genera, including 14 MtGs had rearranged protein-coding genes. This universal primers combined with LR-HY is an efficient way to enrich complete MtGs across the entire Amphibia.

INTRODUCTION

Mitochondrial genomes (MtG) from Amphibia have been used to investigate gene rearrangement and duplication for more than a decade (1-5). Increasing the number of sequenced MtGs enables researchers to discover more structural variation, including *ND6* rearrangement (1), *ND5* duplication (2) or rearrangement (6), tandem repeats (3), pseudogene (4), WANCY-tRNA (7), LTPF, and IQM (8) in frogs and salamanders. These elements that are discovered as highly mutated, rearranged, or duplicated are found in control regions, O_L region and other non-coding regions with no fixed pattern (1-8). To understand the structural variation, we need more MtGs for intra/inter-specific comparison (5). However, the methodology for recovering high diversity MtGs is currently limited.

Recently, various methods have been modified or proposed to couple with high-throughput sequencing (HTS, 9-12). There are three main methods: hybridization, long-range PCR (LR-PCR), and genome skimming. Although capture hybridization and LR-PCR are relatively older than genome skimming, the former two methods are applied more often than genome skimming for genome structure variation (13), population genetics (14) and other evolutionary research in non-model species (10,15). LR-PCR can produce good quality sequences, demonstrates coverage evenness (11,14,16), and avoids nuclear copies of mitochondrial genes (14,17,18). Capture hybridization is considered very time-effective for enriching a massive amount of loci distributed separately in animal genome (10). Capture hybridization is also cost-effective by using PCR amplicon to make probes applicable to non-model organisms (9,19). For genome skimming, some high copy number genes, including entire MtGs, *18s*, and *28s rRNA*, could be filtered by computational methods (12,20). However, control regions in the five

bioRxiv preprint doi: https://doi.org/10.1101/080689; this version posted October 14, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Anuran species were not recovered (5), which means there was a loss of information about structural variation. Data produced by genome skimming includes a large amount of low quality sequences that needs to be removed prior to assembly (5). Moreover, its potential application to other complex structural, variable genomic loci that do not have a high copy number is limited when compared to LR-PCR and capture hybridization (13,21). However, LR-PCR and capture hybridization are also not ideal. On the one hand, degraded DNA effects LR-PCR and success appears to be stochastic when applied to wide taxonomic sample such as in Arthropoda: Araneae (22). On the other hand, capture hybridization limited to distance capture within 20% between probe and target DNA such as in Mammalia (23,24).

To address limitations in methodology for obtaining highly variable loci and rearranged genes in MtGs, we propose an improved *de novo* pipeline by combining LR-PCR and capture hybridization using a new set of primers and a hybridizing long-range library (LR-HY). Our aim is to produce wide taxonomic MtG amplicons with a set of universal primers, and effectively capture highly variable loci using LR-HY. We anticipate that the experimental outcomes will help evolutionary biologists to gain insight to the relationship between gene rearrangement and evolution.

MATERIAL AND METHODS

The general pipeline of this study is shown in Figure 1A. All the samples used in this study are from Amphibia of all the three orders, 12 families, 54 genera and 106 species (Supplementary Table S1) from China and Southeast Asia. For highly diverse taxa, we chose at least one species representative of the genera, and identified the species based on morphology and confirmed with barcode gene *CO1* or *CYTB* (sample information in Supplementary Information S1; method in Supplementary Information S2). DNA samples were extracted using the phenol-chloroform method (25), precipitated with 100% isopropanol and purified with 80% ethanol. Concentrations of DNA samples were measured with a NanoDrop 1000 Spectrophotometer (Thermo). We checked the degree of DNA samples degradation with 0.8% agarose gel and categorized the degree of degradation into four types: no or minor degradation, medium degradation, complete degradation and low concentration (Supplementary Information S1: DNA quality).

bioRxiv preprint doi: https://doi.org/10.1101/080689; this version posted October 14, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Four forward and four reverse primers were designed (Table 1). To achieve universality and avoid impact on the gene rearrangement, we designed degenerate primers on the conservative regions as shown in table 1. Primer pair F1/R2 was used to amplify OA1 (expected length: >14 kb, Figure 1B, green), which covered all the protein coding and control regions. Primer pair F2/R4 was used to amplify OS1 (expected length: >2 kb, figure 1B, green) covered two rRNA genes: *16s rRNA* and a portion of *12s rRNA*. For medium and complete degraded samples (Supplementary Information S1: DNA quality), we designed primers to obtain two similar length amplicons. TF1 was amplified using (F2/F4)/R3 (expected length: 5–9 kb, figure 1B, red) and TR1 was amplified using F3/(R1/R2) (expected length: 7–12 kb, Figure 1B, red). We termed OA1 and OS1 amplification as OA1/OS1 and the alternative pair of amplicons, TF1 and TR1, as TF1/TR1 (Details of the two strategies applied to all of the samples are listed in Supplementary Information S1: Enrichment method).

All of the primer structures were delicately refined. We separated the primers into two regions: the 5' non-degenerate clamp region and the 3' degenerate core region (26). The 3' degenerate core region contained almost all the degenerate points for increasing the possibility mapped to the template. To stabilize the extension of the polymerase, we increased the GC content of the 5' non-degenerate clamp and the AT content of at the beginning of the 3' degenerate core region (27).

Each LR-PCR was conducted in 25 µL reactions containing 50–200 ng template, 5 µL 5× PCR buffer, 3 µL 2.5 mM dNTP (Takara), 0.8 µL 10 µM forward and reverse primers (Invitrogen), and 1 µL LongAmp DNA Polymerase (New England BioLabs, NEB). We used a thermal cycler (Applied Biosystems 2720, 9700 or Veriti) for LR-PCR and its conditions were as follows: initial incubation at 95 ℃ for 1 min, 30–32 cycles at 94 ℃ for 10 s, 58 ℃ for 40 s, and 65 ℃ extensive for variable times, a final extension at 65 ℃ for 10 min, and hold 10℃. Extension times were 3 min for OS1, 10 min for TF1 and TR1, and 16 min for OA1 (Figure 1B). To assure high product concentrations for probe making and library construction, we pooled multiple tubes of the LR-PCR products (number range from 3-10 depending on PCR efficiency for these samples; data not collected). The pooled products were gel-purified using a WIZARD gel extraction kit (Promega).

bioRxiv preprint doi: https://doi.org/10.1101/080689; this version posted October 14, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

To make probe for capture hybridization, the probe was automatically generated from LR-PCR amplicons using a BioNick Labeling Kit (Invitrogen) according to the manufacturer's protocol with the slight modification of extending incubation time to 90 min. The ratio of the TF1 to TR1 amplicon was 5:8. The ratio of the OS1 to OA1 amplicon was 1:12 according to amplicon length (and empirically adjusted according to sequence depth). We also pooled 33 total DNA samples from different species to construct one library to save time and expense. The 33-pooled library was captured using a mixed probe set. We chose 26 closely related amplicon pairs and mixed 130 ng of each of them to make a mixed probe set (the 26 pairs of amplicons listed in Supplementary Information S1: hybridization parameters).

The general pipeline of library construction was shearing, end-repair, adaptor ligation, size selection and library enrichment. For the library construction 1, initial DNA quantity was 130 ng. To obtain a 2 kb fragment, we sheared the DNA samples in a Focused-ultrasonicator M220 (Covaris) by selecting the method DNA_2000bp_200_ul_Clear_microTUBE for 12 min for *R. jiemuxiensis, O. zhangyapingi,* and the 33-pooled DNA samples. To obtain 400 bp fragment for standard capture hybridization, we use an IonShear kit (ThermoFisher) to shear for 200 s in an open thermal cycler. End-repair was carried out in 100 µL reactions, containing 130 ng sheared DNA, 20 µL 5 × End Repair Buffer, and 1 µL End Repair Enzyme. Adaptor ligation for 130 ng of sheared DNA was mixed with 1.6 µL (Ion Xpress Barcode Adapter Kits from 1 to 96), 10 µL 10 × Ligase Buffer, 2 µL dNTP Mix, 2 µL DNA Ligase and 8 µL Nick Repair Polymerase (Ion Plus Fragment Library Kit). This mixture was incubated for 20 min at 25 ℃ in a thermal cycler. The temperature was then increased to 72 ℃ incubated for 5 min. Sheared DNA of *R. jiemuxiensis* and *O. zhangyapingi* were selected by Ampure bead (Beckman) with a corresponding volume of 0.4 of the DNA solution (i.e., 10 µL sample of DNA gets 4 µL of Ampure beads) according to the manufactory's protocol to reduce short fragments. Library amplification was carried out in a PCR volume of 50 µL, containing un-enriched library, 10 µL 5× PCR buffer, 5 µL 2.5 mM dNTP, 2 µL of 10 µM forward and reverse primers (Invitrogen), and 2 µL LongAmp DNA Polymerase (NEB). The PCR conditions were as follows: 95 ℃ for 1 min, then 15 cycles of 94 ℃ for 10 s, 58 ℃ for 40 s, 65 ℃ for 3 min, and finally 65 ℃ for 10 min followed by holding at 4 ℃. The reagent usage of the 33-pooled sample was different. We mixed DNAs at the same quantity of 130 ng from the 32 samples and 0.13 ng LR-PCR products of *Ichthyophis bannanicus* as an internal control. The 33-pooled samples was conducted in a reaction of 200

bioRxiv preprint doi: https://doi.org/10.1101/080689; this version posted October 14, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

μL for end repair; the amount of End Repair Buffer and End Repair Enzyme were doubled. In adaptor ligation, the amount of adaptor, Nick Repair Polymerase, and DNA ligase were doubled for the 33-pooled samples. Size-selection used a 1% agarose gel to obtain an approximately 2 kb long fragment.

In the capture hybridization, we mixed 2× hybridization buffer (Agilent), 10× blocking agent, 2 μl Human Cot-1 DNA (Agilent), 2 μl of blocking adaptors (Ion Plus Kit, ThermoFisher) and certain ratio of library and probe; 1:10 for single-sample library to probe and 1:1 for the 33-pooled library to the 26-mixed probe. This mixture was placed in a thermocycler for 5 min at 95 ℃, and then incubated for 72 hr at 65 ℃–58 ℃ while reducing 2 ℃ every 24 hr. Following incubation, samples were washed with streptavidin beads (M-270, Invitrogen) following the protocol described in (27), but with the addition of a one-minute vortex. Amplification was conducted using a Library Amplification Kit (KAPA) with 25 μL HiFi mix, 21 μL selected fragment solution and 4 μL primer mix. The PCR conditions were as follows: 98 ℃ for 1 min, eight cycles of 98 ℃ for 15 s, 58 ℃for 30 s, and 72 ℃ for 1 min, followed by 72 ℃ for 5 min and hold at 4℃.

To fill gaps near *ND4* and *ND5* for 4 samples, *Hylarana taipehensis, Liurana alpinus, Parapelophryne scalpta,* and *Leptobrachium ailaonicum*, we amplified a fragment range from *COX3* to *CYTB* using primer F3 and 5'-GGrATdGAdCGdAGrATdGCrTAnGC-3', with the previously described condition and an extension time of 8 min. Then, we used 130 ng of these amplicons for shearing.

For the library construction 2, LR-PCR amplicons were pooled at ratios of 5:8 for TF1 and TR1 and 1:12 for OS1 and OA1 (In total 130 ng). For LR-PCR product with a low concentration, we purified them again using Ampure beads (Beckman) for shearing with uniform smear patterns. Downstream experiments were followed the protocol in (https://ioncommunity.thermofisher.com/servlet/JiveServlet/downloadBody/3323-102-7-22242/MAN0007044_RevA_UB_3March2014.pdf) with the following modifications. The mixed amplicon libraries were sheared for 200 s using an IonShear kit (ThermoFisher) in an open thermocycler. For the 33-pooled samples, the shearing time was 120s. The conditions of adaptor ligation and amplification were described previously.

bioRxiv preprint doi: https://doi.org/10.1101/080689; this version posted October 14, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

In the sequencing experiment, an Ion Torrent Personal Genome Machine (PGM) was used to sequence because it is fast and relatively inexpensive in terms of each run, not in terms of price per base. Each run using 316 chip generated over 800 Mb for 80 samples and the data size for each sample was more than 10 Mb in general. These generated data are sufficient for *de novo* assembly. Libraries were brought to the same molarity before emulsion PCR according to the following formula:

$$Conc = 1.515 \times C \times 100/\text{Length}$$

C represented the concentration (ng/µl) quantified using Qubit 2.0 (Invitrogen); Length (bp) represented the peak value measured using 2100 Bioanalyzer (Agilent). *Conc* represented molarity (pM). We diluted the molarity of the pooled libraries (300–400 bp insert) to 18–20 pM (instead of 26 pM as recommended in the manufacturer's protocol) to reduce the percentage of polyclones for increasing data output.

Base-calling and quality control were done automatically by Torrent Suite v4.0.2 to generate qualified data without the adaptor sequence. To assess sequence-quality, we canceled quality control in the Torrent Suite to obtain raw data with the adaptor sequence which was subsequently trimmed using AlienTrimmer 0.4.0 (29). We assembled the qualified data using both SPAdes 3.5 (30) and Mira 4.0 (X) to get contigs. Before the MtG was fully assembled, the contig pool was re-assembled using GeneStudio Professional 2.2.0.0 and manually adjusted. Then, we annotated the MtGs using MITOS (31) with default parameters. If there were protein coding regions in control region, we re-annotated them by setting the e-value up to 10^{-5} to verify whether it is pseudogene or not. We used Novocraft 3 (http://www.novocraft.com) or mrsFAST 3.3.0 (31) to obtain mapped reads. Then we use these mapped reads to correct mismatch in coding regions automatically using the mapping model in Mira 4.0 (32). We curated homopolymer error manually by referring to annotation results and aligned files shown in IGV 2.3.46 (33).

RunMapping in Newbler 2.9 was used to generate an AlignmentInfo file and its Total Depth column was used to draw coverage distribution graphs. Reads in the 33-pooled library were assigned to their corresponding species by employing a conservative parameter setting of >98% identity and >95% for region mapping. For the 40 cross-loci (Figure 3), the region mapping was set to >50%. The average read number (Figure 3) was calculated by using mapped read number divided length of the loci. The length of these loci from 12s_1 to CYTB_3 were 346, 343, 323, 427, 400, 419, 437, 324, 324, 324, 350, 348, 351, 390, 390, 390, 390, 350, 338, 178(complete *apt8*), 351, 351, 392, 392, 352, 303, 346, 347, 347, 347, 370,

bioRxiv preprint doi: https://doi.org/10.1101/080689; this version posted October 14, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

371, 371, 370, 382, 269, 270, 370, 383, and 382. DnaSP 5.10 (34) was used to generate slide-window data. Tandem repeats were calculated with TRF v4.07b (35). Kimura 2-parameter (K2P) distances and variation the 40 loci were calculated using MEGA6 (36). Pearson's correlation and linear regression were performed by using R (http://www.R-project.org). Similarity among MtGs was measured using the BLAST function on the National Center for Biotechnology Information (37).

RESULTS

Universal primers, gene rearrangement, and degraded samples

We designed eight general primers and successfully tested them against at least one species in all three orders, including 12 families of amphibians. At the genus level, 92.59% (50/54) of included genera were successfully enriched in at least one sample. The success ratio at the species level was only 74.5% (79/106). Twenty six species were unable to be amplified TR1, yet only one species was unable to be amplified with TF1 (Supplementary Table S1: Enrichment method). TF1/TR1 had a higher success rate than OA1/OS1 in the medium degraded samples. Specifically, TF1/TR1 was recovered in 49 samples, including 11 medium degraded or completely degraded samples (Supplementary Table S1: DNA quality). Moreover, its amplicon have high concentration than OS1/OA1 for making high quality probes.

Highly variable regions and LR-HY

We used a probe of *Rana culaiensis* to capture a closely related mtDNA from *R. jiemuxiensis* (*CO1* K2P = 8.2%). Two gaps still existed in the MtG of *R. jiemuxiensis* at the end of *ND5*, *ND6* and in the non-coding region. These gaps occurred at relatively distant loci of the two MtGs (Figure 2B: black).

We mixed the 26 pairs of amplicons to make probes to capture the 33-pooled sample. The *CO1* K2P distance for target mtDNA to the closest probe range from 0 to 21.8%. In total, 33.19% (23318/70263) of reads mapped to their reference genomes. The correlation analysis shows that there is a significantly negative correlation between variable loci and sequence depth ($P = 3 \times 10^{-5}$, Pearson's correlation). The highest variable regions were ND5_4, ND6_2, apt8, apt6_1 and ND2_3 (Figure 3). These regions have a length range from 178 bp to approximately 2 kb (including lateral non-coding region)

bioRxiv preprint doi: https://doi.org/10.1101/080689; this version posted October 14, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

in these 33 samples. The 33 control region sequences were too variable to align and the lengths were also variable, ranging from 616 bp (*Ichthyophis bannanicus*) to 3,806 bp (*Kurixalus odontotarsus*).

We applied LR-HY to capture a 2 kb library from *Rana jiemuxiensis* and *Onychodactylus zhangyapingi* separately with the same probe made from the *Rana culaiensis* LR-PCR amplicon. As compared to the standard capture hybridization methods using 400 bp library, the LR-HY greatly improved the sequence coverage near *ND5* and *ND6*; only a 400 bp gap in the repetitive region of the MtG of *R. jiemuxiensis* (Figure 2A: green). For the long distance MtG of *O. zhangyapingi*, no gap remained (Figure 2B: green).

The other 27 out of 33 MtGs were also recovery simultaneously. Twelve out of 27 MtGs had small gaps, which may be due to sequence incompletion (Supplementary Information S3). The read number among samples also variable (detailed in the discussion).

Verifying results

All *CO1* genes were sequenced using the Sanger method. The results are identical to the HTS results, except for *Liurana medogensis*, which was unable to be Sanger sequenced. Consensus results among our methods were evaluated. MtGs of *R. jiemuxiensis, O. zhangyapingi, Kurixalus odontotarsus, Occidozyga martensii* and *Babina adenopleura*, were prepared via LR-PCR and the hybridization method. The same results were obtained from all approaches except for a few homopolymer differences.

To check for possible effects of nuclear copies of mitochondrial genes (numts) in assembled MtGs, we translated all protein-coding genes to amino acids. There was no stop codon in the sequences except *ND6* in *Quasipaa yei*. We re-sequenced following the Sanger method to confirm that the two results were identical. To check for possible effects of numts in generated data from LR-PCR amplicons, we distributed the data from 82 samples to their genome, there are 1.76 % (44156/2504790) reads not map to their reference genome. In these unmapped reads pool, there are 41.38% (18273/44156) reads cross-samples contaminated. In total, only approximate 1.03% reads were unmapped to the MtGs.

Rearranged coding gene

Fifteen rearrangement events occurred in a coding gene in this study. Fourteen of the protein-coding genes were recovered (Table 3). These events all were concentrated in *ND5* and *ND6* in the four families. Rearrangement events occurred in control regions near *ND6* in *Kalophrynus interlineatus*, *ND5*

bioRxiv preprint doi: https://doi.org/10.1101/080689; this version posted October 14, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

in 10 Rhacophoridae species, two in Dicroglossidae species, and one Occidozygidae species. In another Rhacophoridae species *Buergeria oxycephala*, the *ND5* inserted between *16s rRNA* and *ND1*, which has not previously been reported.

DISCUSSION

According to the variation pattern in amphibian MtGs, sequencing a length of 2–3 kb is suitable for enrichment of high variable loci. It is possible that a fragment length of >3 kb could obtain longer target DNA and its lateral regions, but it is not recommended to exceed > 10 kb, because extremely high quality DNA samples are required. For medium degraded or low concentration samples, we adjusted the use of Ampure beads to remove short fragments.

We also observed that the capture ability of the home-made probe was not limited to a fixed threshold. For example, sequences between *R. culaiensis* and *R. jiemuxiensis* differed by approximately 15% in the gap between *ND5* and *ND6*. This variation was much smaller than the K2P of 25.5% for the *CO1* between *R. culaiensis* and *O. zhaoermii*, which had relatively high sequence depth. Actually, K2P between *R. jiemuxiensis* and *O. zhaoermii* is larger than 15% cross almost the regions, except the most conservative region in *16s rRNA* (Figure 2A and B: black line between the blue dashed lines). This indicated that the capture ability of the probe depended to some degree on the variation of a gene region.

For coding gene rearrangement, almost all the tree frog species (Rhacophoridae) had a rearranged *ND5* adjacent to or within a control region, except *Buergeria oxycephala,* which had it inserted into another position between *16s rRNA* and *ND1*. Few species of frog and salamander had rearranged *ND6* adjacent to control region and the entire avian class fixed this gene rearrangement. Alam et al. discovered *ND5* duplication: the two identical *ND5*s in the control region of *Hoplobatrachus tigerinus* (NC_014581) and two *ND5*s with 83.5% similarity in *Trichobatrachus robustus* (NC_023382, 2). In addition to coding gene rearrangement and duplication, when we annotated the samples, we observed 16 relic regions from different species, such as *ND5* pseudogene found in *Quasipaa spinose* and *CYTB* pseduogenes in *Andrias davidianus*, *Echinotriton chinhaiensis* and *Quasipaa shini*. Moreover, those duplicated gene, pseudogenes and rearranged loci always follow tandem repeats with a length varying

bioRxiv preprint doi: https://doi.org/10.1101/080689; this version posted October 14, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

from tens to thousands in both inter/intra-species. This potentially indicated emergence or disappearance of a gene due to gene duplication.

The reads number cross-samples is extremely variable in the 33-pooled library. For example, we selected 33 samples from different species for capture hybridization. Read-number varied from one sample to another (Figure 3), and in some cases the difference was substantial. *Bombina orientalis* only had two reads while *Limnonectes bannaensis* had 19267 reads. Linear regression analysis could not establish an association between similarity of probe-target DNA and the number of reads (P = 0.99, linear regression). Then we sequenced a mixed sample of six total DNA samples using a shotgun sequencing method without capture hybridization (Table2). The reads number cross-sample was significantly correlated with the result of standard capture hybridization and not significantly correlated for LR-HY (P = 0.011 and P = 0.074 for standard capture hybridization and LR-HY respectively, Pearson's correlation). Hawkins et al. used qPCR to check whether or not the mtDNA enriched using the probe (24). We also recommend to check the mtDNA concentration before sequencing for those low concentration samples. We could separate them from other high concentration samples for capture hybridization.

We found that the most conservative region is very suitable to be used to design universal primers. The relatively conservative regions in the MtGs are *12s rRNA*, *16s rRNA*, *COX1*, *COX2*, and *COX3* (Figure 3). Seven out of eight of our primers were designed in these regions (Table 1). The conservative regions and variable regions are cross-distributed in the two rRNA genes. For example, *16s rRNA* could be divided into five regions according to the degree of conservation: i, ii, iii, iv, and v (Figure 2B: black line shows the degree of conservation and the five regions labeled in blue). The three conservative regions, i, iii, and v, are intercepted by the variable regions, ii and iv. For the coding gene, we observed that the conservative regions are the first and second codons. The third codon is always variable and require design degenerate points in the primers F3 and R3.

We successfully obtained MtG amplicons from Amphibia. We also extended the application of these primers to other avian and mammalian species, such as gibbons (in press). The probes successfully captured complete MtG of different species using DNA extracted from stool in which the DNA quality was considered medium or highly degraded. Additionally, we have already applied the primers to hundreds of mammal samples (data no shown), including Eulipotyphla, Primatesa, Rodentia, Chiroptera,

bioRxiv preprint doi: https://doi.org/10.1101/080689; this version posted October 14, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Carnivora, Perissodactyla, and Artiodactyla. Twenty-three avian samples were tested and were successfully amplified with primer pair F3/(R1/R2) (data no shown). Therefore, we recommend our primers for application on amniotic samples.

ACCESSION NUMBERS

High-throughput sequencing data have been deposited in the SRA under the accession numbers SRP090718 and in the GenBank: KX021903-KX022007, KX147643, and KX147644.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online: Supplementary Table S1-S2, Supplementary Information S2-S4.

ACKNOWLEDGEMENT

We thank Jing Che's research group for specimen collection and identification, and especially Hong-man Chen who examined species information and its *CO1* sequence. We thank Dong Wang, Ya-han Yang, Kong-Wah Sing, and Elizabeth Georgian for reviewing the manuscript.

FUNDING

This work was supported by the Ministry of Science and Technology of China (MOST no. 2012FY110800 to W.W.) and the National Natural Science Foundation of China (NSFC no. 31090251 to Y.Z.).

Conflict of interest statement. None declared.

REFERENCES

1. Boore, J.L. (1999) Animal mitochondrial genomes. *Nucleic Acids Res.*, **27**, 1767-1780.
2. Alam, M.S., Kurabayashi, A., Hayashi, Y., Sano, N., Khan, M.R., Fujii, T. and Sumida, M. (2010) Complete mitochondrial genomes and novel gene rearrangements in two dicroglossid frogs,

bioRxiv preprint doi: https://doi.org/10.1101/080689; this version posted October 14, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Hoplobatrachus tigerinus and Euphlyctis hexadactylus, from Bangladesh. *Genes Genet. Syst.*, **85**, 219-232.

3. Mueller, R.L. and Boore, J.L. (2005) Molecular mechanisms of extensive mitochondrial gene rearrangement in plethodontid salamanders. *Mol. Biol. Evol.*, **22**, 2104-2112.

4. Xia, Y., Zheng, Y., Miura, I., Wong, P.B., Murphy, R.W. and Zeng, X. (2014) The evolution of mitochondrial genomes in modern frogs (Neobatrachia): nonadaptive evolution of mitochondrial genome reorganization. *Bmc Genomics*, **15**, 691.

5. Machado, D.J., Lyra, M.L. and Grant, T. (2016) Mitogenome assembly from genomic multiplex libraries: comparison of strategies and novel mitogenomes for five species of frogs. *Mol. Ecol. Resour.*, **16**, 686-693.

6. Irisarri, I., San Mauro, D., Abascal, F., Ohler, A., Vences, M. and Zardoya, R. (2012) The origin of modern frogs (Neobatrachia) was accompanied by acceleration in mitochondrial and nuclear substitution rates. *Bmc Genomics*, **13**, 626.

7. San Mauro, D., Gower, D.J., Zardoya, R. and Wilkinson, M. (2006) A hotspot of gene order rearrangement by tandem duplication and random loss in the vertebrate mitochondrial genome. *Mol. Biol. Evol.*, **23**, 227-234.

8. Zhang, P., Liang, D., Mao, R.L., Hillis, D.M., Wake, D.B. and Cannatella, D.C. (2013) Efficient sequencing of Anuran mtDNAs and a mitogenomic exploration of the phylogeny and evolution of frogs. *Mol. Biol. Evol.*, **30**, 1899-1915.

9. Bekaert, B., Ellerington, R., Van den Abbeele, L. and Decorte, R. (2016) In-Solution Hybridization for the Targeted Enrichment of the Whole Mitochondrial Genome. *Methods Mol. Biol.*, **1420**, 173-183.

10. Gasc, C., Peyretaillade, E. and Peyret, P. (2016) Sequence capture by hybridization to explore modern and ancient genomic diversity in model and nonmodel organisms. *Nucleic Acids Res.*, **44**, 4504-4518.

11. Jia, H., Guo, Y., Zhao, W. and Wang, K. (2014) Long-range PCR in next-generation sequencing: comparison of six enzymes and evaluation on the MiSeq sequencer. *Sci. Rep.*, **4**, 5737.

bioRxiv preprint doi: https://doi.org/10.1101/080689; this version posted October 14, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

12. Dodsworth, S. (2015) Genome skimming for next-generation biodiversity analysis. *Trends Plant Sci.*, **20**, 525-527.

13. Carvalho, C.M., Pehlivan, D., Ramocki, M.B., Fang, P., Alleva, B., Franco, L.M., Belmont, J.W., Hastings, P.J. and Lupski, J.R. (2013) Replicative mechanisms for CNV formation are error prone. *Nat. Genet.*, **45**, 1319-1326.

14. Cui, H., Li, F.Y., Chen, D., Wang, G.L., Truong, C.K., Enns, G.M., Graham, B., Milone, M., Landsverk, M.L., Wang, J. *et al.* (2013) Comprehensive next-generation sequence analyses of the entire mitochondrial genome reveal new insights into the molecular diagnosis of mitochondrial DNA disorders. *Genet. Med.*, **15**, 388-394.

15. Paijmans, J.L., Fickel, J., Courtiol, A., Hofreiter, M. and Forster, D.W. (2016) Impact of enrichment conditions on cross-species capture of fresh and degraded DNA. *Mol. Ecol. Resour.*, **16**, 42-55.

16. Harismendy, O. and Frazer, K.A. (2009) Method for improving sequence coverage uniformity of targeted genomic intervals amplified by LR-PCR using Illumina GA sequencing-by-synthesis technology. *Biotechniques*, **46**, 229-231.

17. Li, M., Schroeder, R., Ko, A. and Stoneking, M. (2012) Fidelity of capture-enrichment for mtDNA genome sequencing: influence of NUMTs. *Nucleic Acids Res.*, **40**, e137.

18. Gibbons, J.G., Branco, A.T., Yu, S.K. and Lemos, B. (2014) Ribosomal DNA copy number is coupled with gene expression variation and mitochondrial abundance in humans. *Nat. Commun.*, **5**, 4850.

19. Penalba, J.V., Smith, L.L., Tonione, M.A., Sass, C., Hykin, S.M., Skipwith, P.L., McGuire, J.A., Bowie, R.C.K. and Moritz, C. (2014) Sequence capture using PCR-generated probes: a cost-effective method of targeted high-throughput sequencing for nonmodel organisms. *Mol. Ecol. Resour.*, **14**, 1000-1010.

20. Richter, S., Schwarz, F., Hering, L., Boggemann, M. and Bleidorn, C. (2015) The Utility of Genome Skimming for Phylogenomic Analyses as Demonstrated for Glycerid Relationships (Annelida, Glyceridae). *Genome Biol. Evol.*, **7**, 3443-3462.

bioRxiv preprint doi: https://doi.org/10.1101/080689; this version posted October 14, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

21. Carvalho, C.M. and Lupski, J.R. (2016) Mechanisms underlying structural variant formation in genomic disorders. *Nat. Rev. Genet.*, **17**, 224-238.

22. Briscoe, A.G., Goodacre, S., Masta, S.E., Taylor, M.I., Arnedo, M.A., Penney, D., Kenny, J. and Creer, S. (2013) Can Long-Range PCR Be Used to Amplify Genetically Divergent Mitochondrial Genomes for Comparative Phylogenetics? A Case Study within Spiders (Arthropoda: Araneae). *Plos One*, **8**, e62404.

23. Mason, V.C., Li, G., Helgen, K.M. and Murphy, W.J. (2011) Efficient cross-species capture hybridization and next-generation sequencing of mitochondrial genomes from noninvasively sampled museum specimens. *Genome Res.*, **21**, 1695-1704.

24. Hawkins, M.T., Hofman, C.A., Callicrate, T., McDonough, M.M., Tsuchiya, M.T., Gutierrez, E.E., Helgen, K.M. and Maldonado, J.E. (2016) In-solution hybridization for mammalian mitogenome enrichment: pros, cons and challenges associated with multiplexing degraded DNA. *Mol. Ecol. Resour.*, **16**, 1173-1188.

25. Sambrook, J. and Russell, D.W. (2006) Purification of nucleic acids by extraction with phenol:chloroform. *CSH Protoc.*, **2006**.

26. Rose, T.M., Henikoff, J.G. and Henikoff, S. (2003) CODEHOP (COnsensus-DEgenerate Hybrid Oligonucleotide Primer) PCR primer design. *Nucleic Acids Res.*, **31**, 3763-3766.

27. Rychlik, W. (1995) Selection of primers for polymerase chain reaction. *Mol. Biotechnol.*, **3**, 129–134.

28. Horn, S. (2012) Target enrichment via DNA hybridization capture. *Methods Mol. Biol.*, **840**, 177-188.

29. Criscuolo, A. and Brisse, S. (2013) AlienTrimmer: a tool to quickly and accurately trim off multiple short contaminant sequences from high-throughput sequencing reads. *Genomics*, **102**, 500-506.

30. Bankevich, A., Nurk, S., Antipov, D., Gurevich, A.A., Dvorkin, M., Kulikov, A.S., Lesin, V.M., Nikolenko, S.I., Pham, S., Prjibelski, A.D. *et al.* (2012) SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.*, **19**, 455-477.

31. Hach, F., Hormozdiari, F., Alkan, C., Hormozdiari, F., Birol, I., Eichler, E.E. and Sahinalp, S.C. (2010) mrsFAST: a cache-oblivious algorithm for short-read mapping. *Nat. Methods*, **7**, 576-577.

bioRxiv preprint doi: https://doi.org/10.1101/080689; this version posted October 14, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

32. Burlibasa, C., Vasiliu, D. and Vasiliu, M. (1999) Genome sequence assembly using trace signals and additional sequence information. *German Conference on Bioinformatics*, **99**, 45–56.

33. Thorvaldsdottir, H., Robinson, J.T. and Mesirov, J.P. (2013) Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief Bioinform.*, **14**, 178-192.

34. Librado, P. and Rozas, J. (2009) DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. *Bioinformatics*, **25**, 1451-1452.

35. Benson, G. (1999) Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.*, **27**, 573-580.

36. Tamura, K., Stecher, G., Peterson, D., Filipski, A. and Kumar, S. (2013) MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. *Mol. Biol. Evol.*, **30**, 2725-2729.

37. Johnson, M., Zaretskaya, I., Raytselis, Y., Merezhuk, Y., McGinnis, S. and Madden, T.L. (2008) NCBI BLAST: a better web interface. *Nucleic Acids Res.*, **36**, W5-9.

TABLE AND FIGURES LEGENDS

Table 1. Primer information.

Primer Name	Sequences*	Location	ID
MtG_16s_1645_F	CAGGCCGGAGCAATCCAGGTCr**GTTTCTA**	*16s rRNA*	F1
MtG_16s_1075_R	AGAGGACArGTGATTryGC**TACCTT**	*16s rRNA*	R1
MtG_12s_600_R	GGACACCGCCAAGTCCTTTGGG**TTTTAA**	*12s rRNA*	R2
MtG_12s_480_F	GCTAGGAAACAAACTGGGAT**TAGATACC**	*12s rRNA*	F2
MtG_cox3_R	AGCTGCGGCTTCAAAkCCr**AArTGrTG**	*COX3*	R3
MtG_cox3_F	ATGGCACACCAAGCACAyGChTwyC**AyATAGT**	*COX3*	F3
MtG_12s_270_F	TCGTGCCAGCCACCGCGG**TTAnAC**	*12s rRNA*	F4
MtG_ND1_R	GAGTCCGTCdGCnAndGGTTG	*ND1*	R4

* Underlined denotes the 3' degenerate core regions. Bolding denotes high AT content.

Table 2. Number of reads according to species in the direct sequence and hybridization libraries.

Species	Reads number		
	Direct sequence library	Standard capture hybridization	LR-HY
Paramesotriton hongkongensis	8	66	153
Bufo tibetanus	73	133	1434
Kaloula borealis	53	439	5011

bioRxiv preprint doi: https://doi.org/10.1101/080689; this version posted October 14, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Babina adenopleura	206	681	5181
Leptobrachium liui	248	473	6993
Kurixalus odontotarsus	405	1317	6354

437

438 **Table 3.** Coding gene rearrangements identified in this study

Species	Rearranged gene	Enrichment method[3]	Family
Fejervarya multistriata	*ND5*	LR-PCR	Dicroglossidae
Fejervarya kawamurai	*ND5*	LR-HY/LR-PCR	Dicroglossidae
Occidozyga martensii	*ND5*	LR-HY/LR-PCR	Occidozygidae
Buergeria oxycephala	*ND5*	LR-PCR	Rhacophoridae
Theloderma rhododiscus	*ND5*[1]	LR-PCR	Rhacophoridae
Polypedates megacephalus	*ND5*	LR-PCR	Rhacophoridae
Feihyla vittatus	*ND5*	LR-HY	Rhacophoridae
Rhacophorus bipunctatus	*ND5*	LR-HY	Rhacophoridae
Gracixalus jinxiuensis	*ND5*	LR-PCR	Rhacophoridae
Kurixalus verrucosus	*ND5*	LR-PCR	Rhacophoridae
Raorchestes longchuanensis	*ND5*[2]	LR-HY	Rhacophoridae
Kurixalus odontotarsus	*ND5*	LR-HY/LR-PCR	Rhacophoridae
Rhacophorus kio	*ND5*	LR-PCR	Rhacophoridae
Rhacophorus translineatus	*ND5*	LR-HY	Rhacophoridae
Kalophrynus interlineatus	*ND6*	LR-PCR	Microhylidae

439 1. This *ND5* inserted between *12s rRNA* and *ND1*; 2. *ND5* of *Raorchestes longchuanensis* failed; 3. The
440 LR-HY in this column was conducted in the 33-pooled samples.
441
442

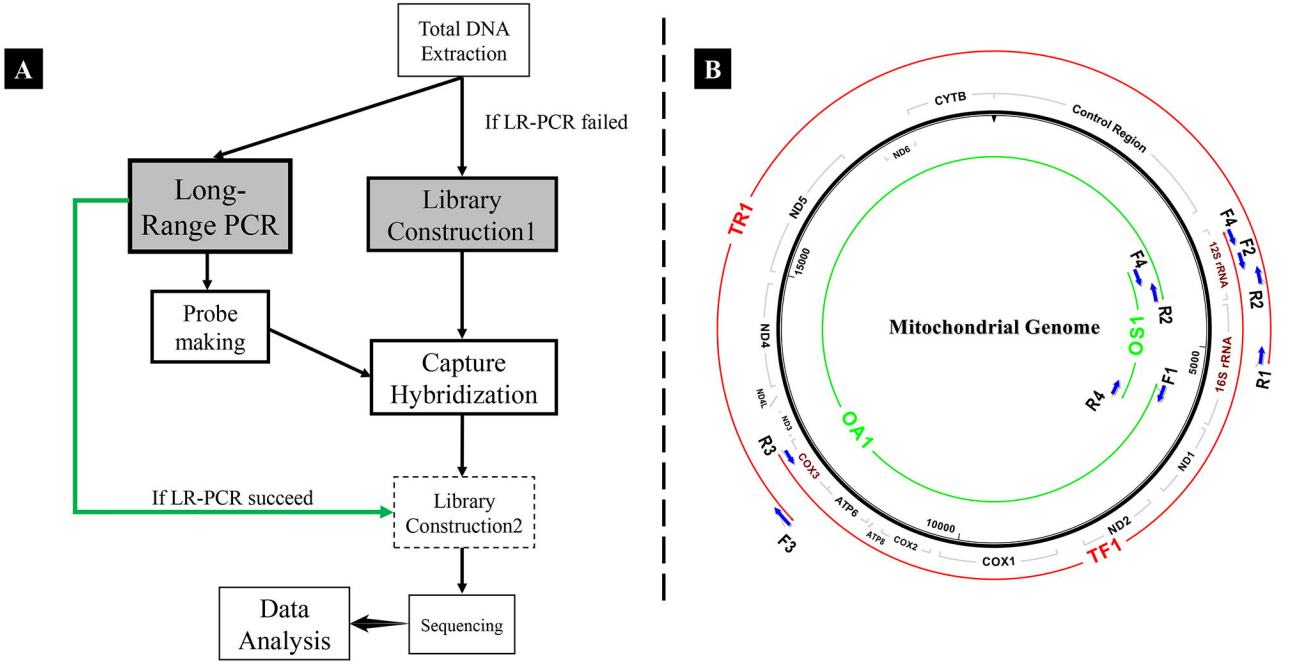
443 **Figure 1**. **A.** Schematic pipeline for enriching mitochondrial DNA to high-throughput sequencing. The
444 green line represents using the pair of LR-PCR amplicons to directly construct a library. Compared to
445 standard library, the LR-HY has modification in library construction one and two. LR-HY requires a long
446 fragment at library construction one and library construction two for PGM sequencing. For standard
447 hybridization, there is no construction library two and sequencing enriched fragments directly; **B**. Two
448 strategies for amplifying MtG, termed OA1/OS1 and TF1/TR1. OA1/OS1: amplification of OA1 and OS1
449 regions uses primers F1/R2 and F4/R4, respectively. TF1/TR1: amplification of fragments TF1 and TR1
450 using primers (F2/F4)/R3 and F3/(R1/R2), respectively.

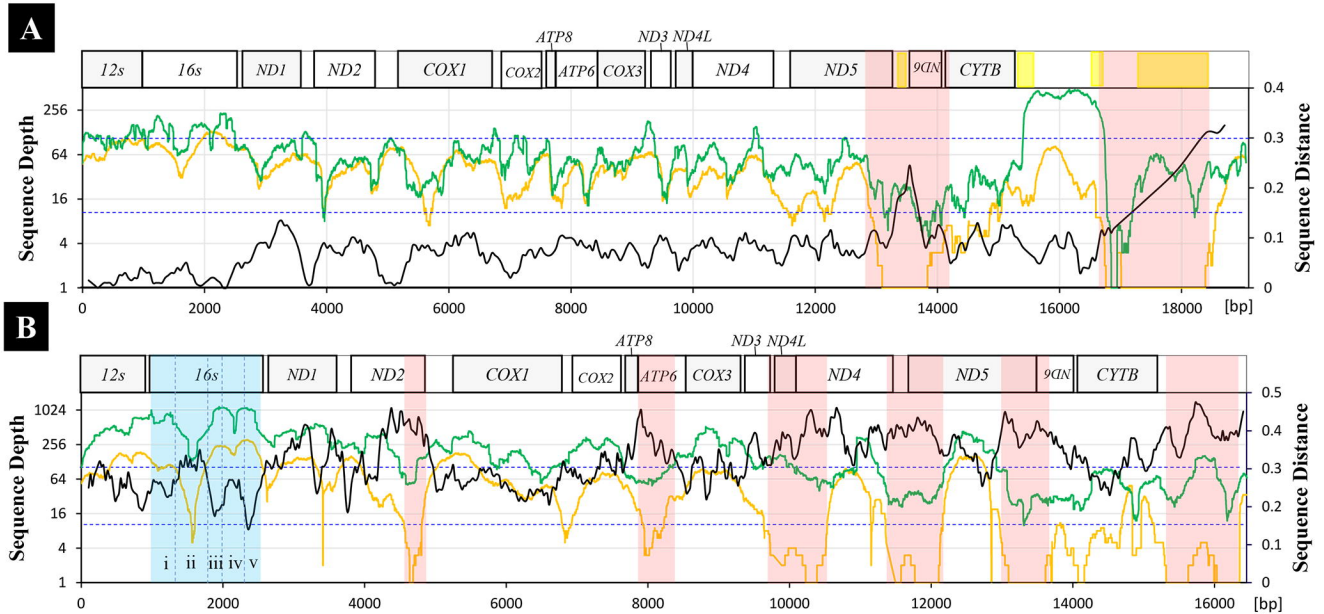
451 **Figure 2**. Coverage distributions for 400 bp and 2 kb library. **A** represents *Rana jiemuxiensis* results by
452 using standard capture hybridization (orange line) and LR-HY (green line). Black line represents DNA

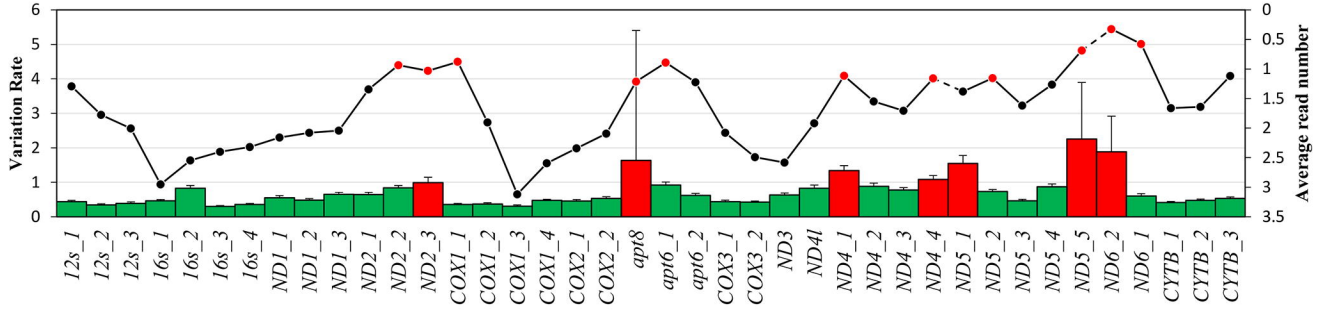
bioRxiv preprint doi: https://doi.org/10.1101/080689; this version posted October 14, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

sequence distance between *R. culaiensis* and *R. jiemuxiensis*. The sliding window length is 50 bp and the step length is 5 bp (below is the same). Dashed lines in A and B are constant at 0.15 and 0.3 sequence distance. The repetitive regions in *R. jiemuxiensis* which is labeled with yellow ranged from 13,424 to 13,572 bp, 15,402 to 15,660 bp, 16,593 to 16,770 bp and 17,382 to 18,498 bp. **B** represents *Onychodactylus zhangyapingi* results by using standard capture hybridization (orange line) and LR-HY (green line). Black line represents DNA sequence distance between *R. culaiensis* and *O. zhangyapingi*. Dashed lines in A and B are constant at 0.15 and 0.3 of K2P. The regions with greatest sequence depth improvement are highlighted with red box. The five regions, i, ii, iii, iv, and v, with different sequence variation in *16s rRNA* are highlighted with blue box.

Figure 3. Variation rate and average reads number cross-region in two rRNA and 13 protein coding genes. The histogram represents variation rate across 40 loci in two rRNA and 13 coding genes. Line plots represent average read number for each loci. Two of the dashed line represents the occurring of *ND5* gene rearrangement in seven species (Table 3).


A
Total DNA Extraction
If LR-PCR failed
Long-Range PCR
Library Construction1
Probe making
Capture Hybridization
If LR-PCR succeed
Library Construction2
Data Analysis
Sequencing
B
Mitochondrial Genome
CYTB
Control Region
ND6
ND5
15000
ND4
ND4L
ND3
COX3
ATP6
ATP8
COX2
10000
COX1
ND2
ND1
5000
16S rRNA
12S rRNA
TR1
TF1
OS1
OA1
F1
F2
F3
F4
R1
R2
R3
R4


A
12s
16s
ND1
ND2
COX1
COX2
ATP8
ATP6
COX3
ND3
ND4L
ND4
ND5
ND6
CYTB
Sequence Depth
Sequence Distance
[bp]
B
i
ii
iii
iv
v


Variation Rate
Average read number
6
5
4
3
2
1
0
0
0.5
1
1.5
2
2.5
3
3.5
12s_1
12s_2
12s_3
16s_1
16s_2
16s_3
16s_4
ND1_1
ND1_2
ND1_3
ND2_1
ND2_2
ND2_3
COX1_1
COX1_2
COX1_3
COX1_4
COX2_1
COX2_2
apt8
apt6_1
apt6_2
COX3_1
COX3_2
ND3
ND4l
ND4_1
ND4_2
ND4_3
ND4_4
ND5_1
ND5_2
ND5_3
ND5_4
ND5_5
ND6_2
ND6_1
CYTB_1
CYTB_2
CYTB_3