

# 1 An improved *de novo* pipeline for enrichment of high diversity 2 mitochondrial genomes from Amphibia to high-throughput 3 sequencing

4

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

7

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

24

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

28 †These authors contributed equally to this work.

29

## 30 **ABSTRACT**

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

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

45

## 46 INTRODUCTION

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

55 Recently, various methods have been modified or proposed to couple with high-throughput  
56 sequencing (HTS, 9-12). There are three main methods: hybridization, long-range PCR (LR-PCR), and  
57 genome skimming. Although capture hybridization and LR-PCR are relatively older than genome  
58 skimming, the former two methods are applied more often than genome skimming for genome structure  
59 variation (13), population genetics (14) and other evolutionary research in non-model species (10,15). LR-  
60 PCR can produce good quality sequences, demonstrates coverage evenness (11,14,16), and avoids  
61 nuclear copies of mitochondrial genes (14,17,18). Capture hybridization is considered very time-effective  
62 for enriching a massive amount of loci distributed separately in animal genome (10). Capture  
63 hybridization is also cost-effective by using PCR amplicon to make probes applicable to non-model  
64 organisms (9,19). For genome skimming, some high copy number genes, including entire MtGs, *18s*, and  
65 *28s rRNA*, could be filtered by computational methods (12,20). However, control regions in the five

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

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

## 80 MATERIAL AND METHODS

81 The general pipeline of this study is shown in Figure 1A. All the samples used in this study are  
82 from Amphibia of all the three orders, 12 families, 54 genera and 106 species (Supplementary Table S1)  
83 from China and Southeast Asia. For highly diverse taxa, we chose at least one species representative of  
84 the genera, and identified the species based on morphology and confirmed with barcode gene *CO1* or  
85 *CYTB* (sample information in Supplementary Information S1; method in Supplementary Information S2).  
86 DNA samples were extracted using the phenol-chloroform method (25), precipitated with 100%  
87 isopropanol and purified with 80% ethanol. Concentrations of DNA samples were measured with a  
88 NanoDrop 1000 Spectrophotometer (Thermo). We checked the degree of DNA samples degradation with  
89 0.8% agarose gel and categorized the degree of degradation into four types: no or minor degradation,  
90 medium degradation, complete degradation and low concentration (Supplementary Information S1: DNA  
91 quality).

92 Four forward and four reverse primers were designed (Table 1). To achieve universality and  
93 avoid impact on the gene rearrangement, we designed degenerate primers on the conservative regions  
94 as shown in table 1. Primer pair F1/R2 was used to amplify OA1 (expected length: >14 kb, Figure 1B,  
95 green), which covered all the protein coding and control regions. Primer pair F2/R4 was used to amplify  
96 OS1 (expected length: >2 kb, figure 1B, green) covered two rRNA genes: *16s rRNA* and a portion of *12s*  
97 *rRNA*. For medium and complete degraded samples (Supplementary Information S1: DNA quality), we  
98 designed primers to obtain two similar length amplicons. TF1 was amplified using (F2/F4)/R3 (expected  
99 length: 5–9 kb, figure 1B, red) and TR1 was amplified using F3/(R1/R2) (expected length: 7–12 kb, Figure  
100 1B, red). We termed OA1 and OS1 amplification as OA1/OS1 and the alternative pair of amplicons, TF1  
101 and TR1, as TF1/TR1 (Details of the two strategies applied to all of the samples are listed in  
102 Supplementary Information S1: Enrichment method).

103 All of the primer structures were delicately refined. We separated the primers into two regions: the  
104 5' non-degenerate clamp region and the 3' degenerate core region (26). The 3' degenerate core region  
105 contained almost all the degenerate points for increasing the possibility mapped to the template. To  
106 stabilize the extension of the polymerase, we increased the GC content of the 5' non-degenerate clamp  
107 and the AT content of at the beginning of the 3' degenerate core region (27).

108 Each LR-PCR was conducted in 25  $\mu$ L reactions containing 50–200 ng template, 5  $\mu$ L 5 $\times$  PCR  
109 buffer, 3  $\mu$ L 2.5 mM dNTP (Takara), 0.8  $\mu$ L 10  $\mu$ M forward and reverse primers (Invitrogen), and 1  $\mu$ L  
110 LongAmp DNA Polymerase (New England BioLabs, NEB). We used a thermal cycler (Applied Biosystems  
111 2720, 9700 or Veriti) for LR-PCR and its conditions were as follows: initial incubation at 95  $^{\circ}$ C for 1 min,  
112 30–32 cycles at 94  $^{\circ}$ C for 10 s, 58  $^{\circ}$ C for 40 s, and 65  $^{\circ}$ C extensive for variable times, a final extension at  
113 65  $^{\circ}$ C for 10 min, and hold 10 $^{\circ}$ C. Extension times were 3 min for OS1, 10 min for TF1 and TR1, and 16  
114 min for OA1 (Figure 1B). To assure high product concentrations for probe making and library  
115 construction, we pooled multiple tubes of the LR-PCR products (number range from 3-10 depending on  
116 PCR efficiency for these samples; data not collected). The pooled products were gel-purified using a  
117 WIZARD gel extraction kit (Promega).

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

126 The general pipeline of library construction was shearing, end-repair, adaptor ligation, size  
127 selection and library enrichment. For the library construction 1, initial DNA quantity was 130 ng. To obtain  
128 a 2 kb fragment, we sheared the DNA samples in a Focused-ultrasonicator M220 (Covaris) by selecting  
129 the method DNA\_2000bp\_200\_ul\_Clear\_microTUBE for 12 min for *R. jiemuxiensis*, *O. zhangyapingi*, and  
130 the 33-pooled DNA samples. To obtain 400 bp fragment for standard capture hybridization, we use an  
131 IonShear kit (ThermoFisher) to shear for 200 s in an open thermal cyclers. End-repair was carried out in  
132 100  $\mu$ L reactions, containing 130 ng sheared DNA, 20  $\mu$ L 5  $\times$  End Repair Buffer, and 1  $\mu$ L End Repair  
133 Enzyme. Adaptor ligation for 130 ng of sheared DNA was mixed with 1.6  $\mu$ L (Ion Xpress Barcode Adapter  
134 Kits from 1 to 96), 10  $\mu$ L 10  $\times$  Ligase Buffer, 2  $\mu$ L dNTP Mix, 2  $\mu$ L DNA Ligase and 8  $\mu$ L Nick Repair  
135 Polymerase (Ion Plus Fragment Library Kit). This mixture was incubated for 20 min at 25  $^{\circ}$ C in a thermal  
136 cyclers. The temperature was then increased to 72  $^{\circ}$ C incubated for 5 min. Sheared DNA of *R. jiemuxiensis*  
137 and *O. zhangyapingi* were selected by Ampure bead (Beckman) with a corresponding volume of 0.4 of  
138 the DNA solution (i.e., 10  $\mu$ L sample of DNA gets 4  $\mu$ L of Ampure beads) according to the manufactory's  
139 protocol to reduce short fragments. Library amplification was carried out in a PCR volume of 50  $\mu$ L,  
140 containing un-enriched library, 10  $\mu$ L 5 $\times$  PCR buffer, 5  $\mu$ L 2.5 mM dNTP, 2  $\mu$ L of 10  $\mu$ M forward and  
141 reverse primers (Invitrogen), and 2  $\mu$ L LongAmp DNA Polymerase (NEB). The PCR conditions were as  
142 follows: 95  $^{\circ}$ C for 1 min, then 15 cycles of 94  $^{\circ}$ C for 10 s, 58  $^{\circ}$ C for 40 s, 65  $^{\circ}$ C for 3 min, and finally 65  $^{\circ}$ C  
143 for 10 min followed by holding at 4  $^{\circ}$ C. The reagent usage of the 33-pooled sample was different. We  
144 mixed DNAs at the same quantity of 130 ng from the 32 samples and 0.13 ng LR-PCR products of  
145 *Ichthyophis bannanicus* as an internal control. The 33-pooled samples was conducted in a reaction of 200

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

149 In the capture hybridization, we mixed 2 $\times$  hybridization buffer (Agilent), 10 $\times$  blocking agent, 2  $\mu$ L  
150 Human Cot-1 DNA (Agilent), 2  $\mu$ L of blocking adaptors (Ion Plus Kit, ThermoFisher) and certain ratio of  
151 library and probe; 1:10 for single-sample library to probe and 1:1 for the 33-pooled library to the 26-mixed  
152 probe. This mixture was placed in a thermocycler for 5 min at 95  $^{\circ}$ C, and then incubated for 72 hr at  
153 65  $^{\circ}$ C–58  $^{\circ}$ C while reducing 2  $^{\circ}$ C every 24 hr. Following incubation, samples were washed with streptavidin  
154 beads (M-270, Invitrogen) following the protocol described in (27), but with the addition of a one-minute  
155 vortex. Amplification was conducted using a Library Amplification Kit (KAPA) with 25  $\mu$ L HiFi mix, 21  $\mu$ L  
156 selected fragment solution and 4  $\mu$ L primer mix. The PCR conditions were as follows: 98  $^{\circ}$ C for 1 min,  
157 eight cycles of 98  $^{\circ}$ C for 15 s, 58  $^{\circ}$ C for 30 s, and 72  $^{\circ}$ C for 1 min, followed by 72  $^{\circ}$ C for 5 min and hold at  
158 4 $^{\circ}$ C.

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

163 For the library construction 2, LR-PCR amplicons were pooled at ratios of 5:8 for TF1 and TR1  
164 and 1:12 for OS1 and OA1 (In total 130 ng). For LR-PCR product with a low concentration, we purified  
165 them again using Ampure beads (Beckman) for shearing with uniform smear patterns. Downstream  
166 experiments were followed the protocol in  
167 ([https://ioncommunity.thermofisher.com/servlet/JiveServlet/downloadBody/3323-102-7-](https://ioncommunity.thermofisher.com/servlet/JiveServlet/downloadBody/3323-102-7-22242/MAN0007044_RevA_UB_3March2014.pdf)  
168 [22242/MAN0007044\\_RevA\\_UB\\_3March2014.pdf](https://ioncommunity.thermofisher.com/servlet/JiveServlet/downloadBody/3323-102-7-22242/MAN0007044_RevA_UB_3March2014.pdf)) with the following modifications. The mixed amplicon  
169 libraries were sheared for 200 s using an IonShear kit (ThermoFisher) in an open thermocycler. For the  
170 33-pooled samples, the shearing time was 120s. The conditions of adaptor ligation and amplification were  
171 described previously.

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

$$177 \text{Conc} = 1.515 \times C \times 100/\text{Length}$$

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

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

193 RunMapping in Newbler 2.9 was used to generate an AlignmentInfo file and its Total Depth column  
194 was used to draw coverage distribution graphs. Reads in the 33-pooled library were assigned to their  
195 corresponding species by employing a conservative parameter setting of >98% identity and >95% for  
196 region mapping. For the 40 cross-loci (Figure 3), the region mapping was set to >50%. The average read  
197 number (Figure 3) was calculated by using mapped read number divided length of the loci. The length of  
198 these loci from 12s\_1 to CYTB\_3 were 346, 343, 323, 427, 400, 419, 437, 324, 324, 324, 350, 348, 351,  
199 390, 390, 390, 390, 350, 338, 178(complete *apt8*), 351, 351, 392, 392, 352, 303, 346, 347, 347, 347, 370,

200 371, 371, 370, 382, 269, 270, 370, 383, and 382. DnaSP 5.10 (34) was used to generate slide-window  
201 data. Tandem repeats were calculated with TRF v4.07b (35). Kimura 2-parameter (K2P) distances and  
202 variation the 40 loci were calculated using MEGA6 (36). Pearson's correlation and linear regression were  
203 performed by using R (<http://www.R-project.org>). Similarity among MtGs was measured using the BLAST  
204 function on the National Center for Biotechnology Information (37).

205

## 206 **RESULTS**

### 207 **Universal primers, gene rearrangement, and degraded samples**

208 We designed eight general primers and successfully tested them against at least one species in  
209 all three orders, including 12 families of amphibians. At the genus level, 92.59% (50/54) of included  
210 genera were successfully enriched in at least one sample. The success ratio at the species level was only  
211 74.5% (79/106). Twenty six species were unable to be amplified TR1, yet only one species was unable  
212 to be amplified with TF1 (Supplementary Table S1: Enrichment method). TF1/TR1 had a higher success  
213 rate than OA1/OS1 in the medium degraded samples. Specifically, TF1/TR1 was recovered in 49  
214 samples, including 11 medium degraded or completely degraded samples (Supplementary Table S1:  
215 DNA quality). Moreover, its amplicon have high concentration than OS1/OA1 for making high quality  
216 probes.

### 217 **Highly variable regions and LR-HY**

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

221 We mixed the 26 pairs of amplicons to make probes to capture the 33-pooled sample. The *CO1*  
222 K2P distance for target mtDNA to the closest probe range from 0 to 21.8%. In total, 33.19%  
223 (23318/70263) of reads mapped to their reference genomes. The correlation analysis shows that there is  
224 a significantly negative correlation between variable loci and sequence depth ( $P = 3 \times 10^{-5}$ , Pearson's  
225 correlation). The highest variable regions were *ND5\_4*, *ND6\_2*, *apt8*, *apt6\_1* and *ND2\_3* (Figure 3).  
226 These regions have a length range from 178 bp to approximately 2 kb (including lateral non-coding region)

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

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

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

### 238 **Verifying results**

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

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

### 251 **Rearranged coding gene**

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

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

258

## 259 **DISCUSSION**

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

265 We also observed that the capture ability of the home-made probe was not limited to a fixed  
266 threshold. For example, sequences between *R. culaiensis* and *R. jiemuxiensis* differed by approximately  
267 15% in the gap between *ND5* and *ND6*. This variation was much smaller than the K2P of 25.5% for the  
268 *CO1* between *R. culaiensis* and *O. zhaoermii*, which had relatively high sequence depth. Actually, K2P  
269 between *R. jiemuxiensis* and *O. zhaoermii* is larger than 15% cross almost the regions, except the most  
270 conservative region in *16s rRNA* (Figure 2A and B: black line between the blue dashed lines). This  
271 indicated that the capture ability of the probe depended to some degree on the variation of a gene region.

272 For coding gene rearrangement, almost all the tree frog species (Rhacophoridae) had a  
273 rearranged *ND5* adjacent to or within a control region, except *Buergeria oxycephala*, which had it inserted  
274 into another position between *16s rRNA* and *ND1*. Few species of frog and salamander had rearranged  
275 *ND6* adjacent to control region and the entire avian class fixed this gene rearrangement. Alam et al.  
276 discovered *ND5* duplication: the two identical *ND5*s in the control region of *Hoplobatrachus tigerinus*  
277 (NC\_014581) and two *ND5*s with 83.5% similarity in *Trichobatrachus robustus* (NC\_023382, 2). In  
278 addition to coding gene rearrangement and duplication, when we annotated the samples, we observed 16  
279 relic regions from different species, such as *ND5* pseudogene found in *Quasipaa spinose* and *CYTB*  
280 pseudogenes in *Andrias davidianus*, *Echinotriton chinhaiensis* and *Quasipaa shini*. Moreover, those  
281 duplicated gene, pseudogenes and rearranged loci always follow tandem repeats with a length varying

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

284 The reads number cross-samples is extremely variable in the 33-pooled library. For example, we  
285 selected 33 samples from different species for capture hybridization. Read-number varied from one  
286 sample to another (Figure 3), and in some cases the difference was substantial. *Bombina orientalis* only  
287 had two reads while *Limnonectes bannaensis* had 19267 reads. Linear regression analysis could not  
288 establish an association between similarity of probe-target DNA and the number of reads ( $P = 0.99$ , linear  
289 regression). Then we sequenced a mixed sample of six total DNA samples using a shotgun sequencing  
290 method without capture hybridization (Table2). The reads number cross-sample was significantly  
291 correlated with the result of standard capture hybridization and not significantly correlated for LR-HY ( $P =$   
292  $0.011$  and  $P = 0.074$  for standard capture hybridization and LR-HY respectively, Pearson's correlation).  
293 Hawkins et al. used qPCR to check whether or not the mtDNA enriched using the probe (24). We also  
294 recommend to check the mtDNA concentration before sequencing for those low concentration samples.  
295 We could separate them from other high concentration samples for capture hybridization.

296 We found that the most conservative region is very suitable to be used to design universal  
297 primers. The relatively conservative regions in the MtGs are *12s rRNA*, *16s rRNA*, *COX1*, *COX2*, and  
298 *COX3* (Figure 3). Seven out of eight of our primers were designed in these regions (Table 1). The  
299 conservative regions and variable regions are cross-distributed in the two rRNA genes. For example, *16s*  
300 *rRNA* could be divided into five regions according to the degree of conservation: i, ii, iii, iv, and v (Figure  
301 2B: black line shows the degree of conservation and the five regions labeled in blue). The three  
302 conservative regions, i, iii, and v, are intercepted by the variable regions, ii and iv. For the coding gene,  
303 we observed that the conservative regions are the first and second codons. The third codon is always  
304 variable and require design degenerate points in the primers F3 and R3.

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

310 Carnivora, Perissodactyla, and Artiodactyla. Twenty-three avian samples were tested and were  
311 successfully amplified with primer pair F3/(R1/R2) (data not shown). Therefore, we recommend our  
312 primers for application on amniotic samples.

313

#### 314 **ACCESSION NUMBERS**

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

317

#### 318 **SUPPLEMENTARY DATA**

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

321

#### 322 **ACKNOWLEDGEMENT**

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

326

#### 327 **FUNDING**

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

330

331 *Conflict of interest statement.* None declared.

332

#### 333 **REFERENCES**

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

- 337 Hoplobatrachus tigerinus and Euphlyctis hexadactylus, from Bangladesh. *Genes Genet. Syst.*, **85**,  
338 219-232.
- 339 3. Mueller, R.L. and Boore, J.L. (2005) Molecular mechanisms of extensive mitochondrial gene  
340 rearrangement in plethodontid salamanders. *Mol. Biol. Evol.*, **22**, 2104-2112.
- 341 4. Xia, Y., Zheng, Y., Miura, I., Wong, P.B., Murphy, R.W. and Zeng, X. (2014) The evolution of  
342 mitochondrial genomes in modern frogs (Neobatrachia): nonadaptive evolution of mitochondrial  
343 genome reorganization. *Bmc Genomics*, **15**, 691.
- 344 5. Machado, D.J., Lyra, M.L. and Grant, T. (2016) Mitogenome assembly from genomic multiplex  
345 libraries: comparison of strategies and novel mitogenomes for five species of frogs. *Mol. Ecol.*  
346 *Resour.*, **16**, 686-693.
- 347 6. Irisarri, I., San Mauro, D., Abascal, F., Ohler, A., Vences, M. and Zardoya, R. (2012) The origin of  
348 modern frogs (Neobatrachia) was accompanied by acceleration in mitochondrial and nuclear  
349 substitution rates. *Bmc Genomics*, **13**, 626.
- 350 7. San Mauro, D., Gower, D.J., Zardoya, R. and Wilkinson, M. (2006) A hotspot of gene order  
351 rearrangement by tandem duplication and random loss in the vertebrate mitochondrial genome.  
352 *Mol. Biol. Evol.*, **23**, 227-234.
- 353 8. Zhang, P., Liang, D., Mao, R.L., Hillis, D.M., Wake, D.B. and Cannatella, D.C. (2013) Efficient  
354 sequencing of Anuran mtDNAs and a mitogenomic exploration of the phylogeny and evolution of  
355 frogs. *Mol. Biol. Evol.*, **30**, 1899-1915.
- 356 9. Bekaert, B., Ellerington, R., Van den Abbeele, L. and Decorte, R. (2016) In-Solution Hybridization  
357 for the Targeted Enrichment of the Whole Mitochondrial Genome. *Methods Mol. Biol.*, **1420**, 173-  
358 183.
- 359 10. Gasc, C., Peyretailade, E. and Peyret, P. (2016) Sequence capture by hybridization to explore  
360 modern and ancient genomic diversity in model and nonmodel organisms. *Nucleic Acids Res.*, **44**,  
361 4504-4518.
- 362 11. Jia, H., Guo, Y., Zhao, W. and Wang, K. (2014) Long-range PCR in next-generation sequencing:  
363 comparison of six enzymes and evaluation on the MiSeq sequencer. *Sci. Rep.*, **4**, 5737.

- 364 12. Dodsworth, S. (2015) Genome skimming for next-generation biodiversity analysis. *Trends Plant*  
365 *Sci.*, **20**, 525-527.
- 366 13. Carvalho, C.M., Pehlivan, D., Ramocki, M.B., Fang, P., Alleva, B., Franco, L.M., Belmont, J.W.,  
367 Hastings, P.J. and Lupski, J.R. (2013) Replicative mechanisms for CNV formation are error prone.  
368 *Nat. Genet.*, **45**, 1319-1326.
- 369 14. Cui, H., Li, F.Y., Chen, D., Wang, G.L., Truong, C.K., Enns, G.M., Graham, B., Milone, M.,  
370 Landsverk, M.L., Wang, J. *et al.* (2013) Comprehensive next-generation sequence analyses of  
371 the entire mitochondrial genome reveal new insights into the molecular diagnosis of mitochondrial  
372 DNA disorders. *Genet. Med.*, **15**, 388-394.
- 373 15. Paijmans, J.L., Fickel, J., Courtiol, A., Hofreiter, M. and Forster, D.W. (2016) Impact of  
374 enrichment conditions on cross-species capture of fresh and degraded DNA. *Mol. Ecol. Resour.*,  
375 **16**, 42-55.
- 376 16. Harismendy, O. and Frazer, K.A. (2009) Method for improving sequence coverage uniformity of  
377 targeted genomic intervals amplified by LR-PCR using Illumina GA sequencing-by-synthesis  
378 technology. *Biotechniques*, **46**, 229-231.
- 379 17. Li, M., Schroeder, R., Ko, A. and Stoneking, M. (2012) Fidelity of capture-enrichment for mtDNA  
380 genome sequencing: influence of NUMTs. *Nucleic Acids Res.*, **40**, e137.
- 381 18. Gibbons, J.G., Branco, A.T., Yu, S.K. and Lemos, B. (2014) Ribosomal DNA copy number is  
382 coupled with gene expression variation and mitochondrial abundance in humans. *Nat. Commun.*,  
383 **5**, 4850.
- 384 19. Penalba, J.V., Smith, L.L., Tonione, M.A., Sass, C., Hykin, S.M., Skipwith, P.L., McGuire, J.A.,  
385 Bowie, R.C.K. and Moritz, C. (2014) Sequence capture using PCR-generated probes: a cost-  
386 effective method of targeted high-throughput sequencing for nonmodel organisms. *Mol. Ecol.*  
387 *Resour.*, **14**, 1000-1010.
- 388 20. Richter, S., Schwarz, F., Hering, L., Boggemann, M. and Bleidorn, C. (2015) The Utility of  
389 Genome Skimming for Phylogenomic Analyses as Demonstrated for Glycerid Relationships  
390 (Annelida, Glyceridae). *Genome Biol. Evol.*, **7**, 3443-3462.

- 391 21. Carvalho, C.M. and Lupski, J.R. (2016) Mechanisms underlying structural variant formation in  
392 genomic disorders. *Nat. Rev. Genet.*, **17**, 224-238.
- 393 22. Briscoe, A.G., Goodacre, S., Masta, S.E., Taylor, M.I., Arnedo, M.A., Penney, D., Kenny, J. and  
394 Creer, S. (2013) Can Long-Range PCR Be Used to Amplify Genetically Divergent Mitochondrial  
395 Genomes for Comparative Phylogenetics? A Case Study within Spiders (Arthropoda: Araneae).  
396 *Plos One*, **8**, e62404.
- 397 23. Mason, V.C., Li, G., Helgen, K.M. and Murphy, W.J. (2011) Efficient cross-species capture  
398 hybridization and next-generation sequencing of mitochondrial genomes from noninvasively  
399 sampled museum specimens. *Genome Res.*, **21**, 1695-1704.
- 400 24. Hawkins, M.T., Hofman, C.A., Callicrate, T., McDonough, M.M., Tsuchiya, M.T., Gutierrez, E.E.,  
401 Helgen, K.M. and Maldonado, J.E. (2016) In-solution hybridization for mammalian mitogenome  
402 enrichment: pros, cons and challenges associated with multiplexing degraded DNA. *Mol. Ecol.*  
403 *Resour.*, **16**, 1173-1188.
- 404 25. Sambrook, J. and Russell, D.W. (2006) Purification of nucleic acids by extraction with  
405 phenol:chloroform. *CSH Protoc.*, **2006**.
- 406 26. Rose, T.M., Henikoff, J.G. and Henikoff, S. (2003) CODEHOP (COnsensus-DEgenerate Hybrid  
407 Oligonucleotide Primer) PCR primer design. *Nucleic Acids Res.*, **31**, 3763-3766.
- 408 27. Rychlik, W. (1995) Selection of primers for polymerase chain reaction. *Mol. Biotechnol.*, **3**, 129–  
409 134.
- 410 28. Horn, S. (2012) Target enrichment via DNA hybridization capture. *Methods Mol. Biol.*, **840**, 177-  
411 188.
- 412 29. Criscuolo, A. and Brisse, S. (2013) AlienTrimmer: a tool to quickly and accurately trim off multiple  
413 short contaminant sequences from high-throughput sequencing reads. *Genomics*, **102**, 500-506.
- 414 30. Bankevich, A., Nurk, S., Antipov, D., Gurevich, A.A., Dvorkin, M., Kulikov, A.S., Lesin, V.M.,  
415 Nikolenko, S.I., Pham, S., Pribelski, A.D. *et al.* (2012) SPAdes: a new genome assembly  
416 algorithm and its applications to single-cell sequencing. *J. Comput. Biol.*, **19**, 455-477.
- 417 31. Hach, F., Hormozdiari, F., Alkan, C., Hormozdiari, F., Birol, I., Eichler, E.E. and Sahinalp, S.C.  
418 (2010) mrsFAST: a cache-oblivious algorithm for short-read mapping. *Nat. Methods*, **7**, 576-577.

- 419 32. Burlibasa, C., Vasiliu, D. and Vasiliu, M. (1999) Genome sequence assembly using trace signals  
 420 and additional sequence information. *German Conference on Bioinformatics*, **99**, 45–56.
- 421 33. Thorvaldsdottir, H., Robinson, J.T. and Mesirov, J.P. (2013) Integrative Genomics Viewer (IGV):  
 422 high-performance genomics data visualization and exploration. *Brief Bioinform.*, **14**, 178-192.
- 423 34. Librado, P. and Rozas, J. (2009) DnaSP v5: a software for comprehensive analysis of DNA  
 424 polymorphism data. *Bioinformatics*, **25**, 1451-1452.
- 425 35. Benson, G. (1999) Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids*  
 426 *Res.*, **27**, 573-580.
- 427 36. Tamura, K., Stecher, G., Peterson, D., Filipski, A. and Kumar, S. (2013) MEGA6: Molecular  
 428 Evolutionary Genetics Analysis version 6.0. *Mol. Biol. Evol.*, **30**, 2725-2729.
- 429 37. Johnson, M., Zaretskaya, I., Raytselis, Y., Merezuk, Y., McGinnis, S. and Madden, T.L. (2008)  
 430 NCBI BLAST: a better web interface. *Nucleic Acids Res.*, **36**, W5-9.

431

## 432 TABLE AND FIGURES LEGENDS

433 **Table 1.** Primer information.

Primer Name	Sequences*	Location	ID
MtG_16s_1645_F	CAGGCCGGAGCAATCCAGGTCr <u>GTTTCTA</u>	16s rRNA	F1
MtG_16s_1075_R	AGAGGACArGTGATTry <u>GCTACCTT</u>	16s rRNA	R1
MtG_12s_600_R	GGACACCGCCAAGTCC <u>TTTGGGTTTTAA</u>	12s rRNA	R2
MtG_12s_480_F	GCTAGGAAACAACTGGGATTAGATACC	12s rRNA	F2
MtG_cox3_R	AGCTGCGGCTTCAA <u>AAkCCrAArTGrTG</u>	COX3	R3
MtG_cox3_F	ATGGCACACCAAGCACAY <u>GChTwyCAyATAGT</u>	COX3	F3
MtG_12s_270_F	TCGTGCCAGCCACCGCGGTT <u>AnAC</u>	12s rRNA	F4
MtG_ND1_R	GAGTCCGTCDGCnAndGGTTG	ND1	R4

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

435

436 **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
<i>Paramesotriton hongkongensis</i>	8	66	153
<i>Bufo tibetanus</i>	73	133	1434
<i>Kaloula borealis</i>	53	439	5011

<i>Babina adenopleura</i>	206	681	5181
<i>Leptobrachium liui</i>	248	473	6993
<i>Kurixalus odontotarsus</i>	405	1317	6354

437

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

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

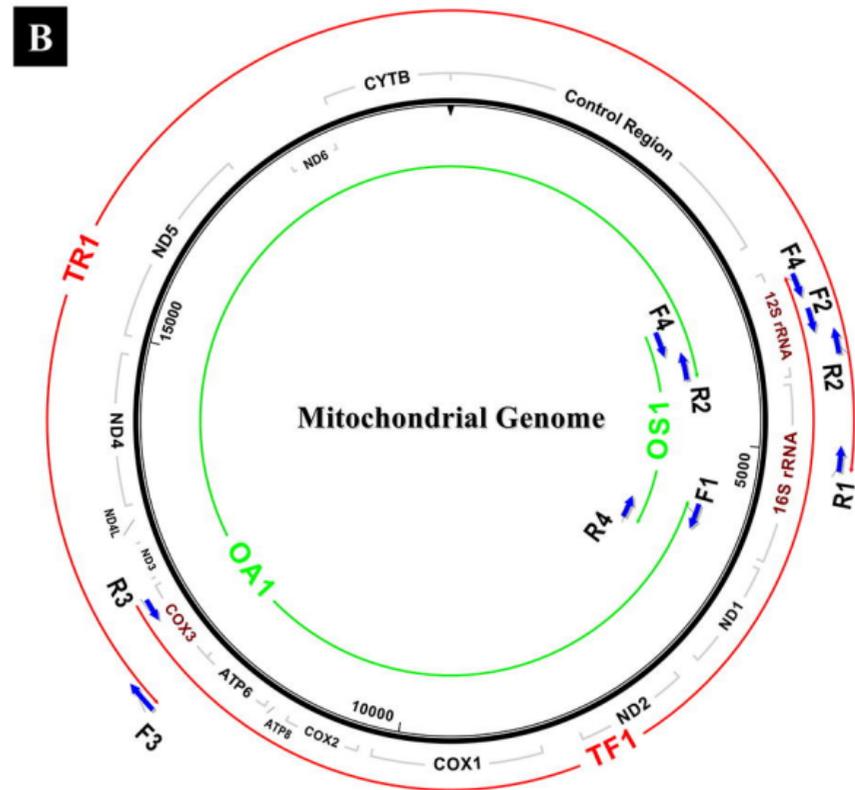
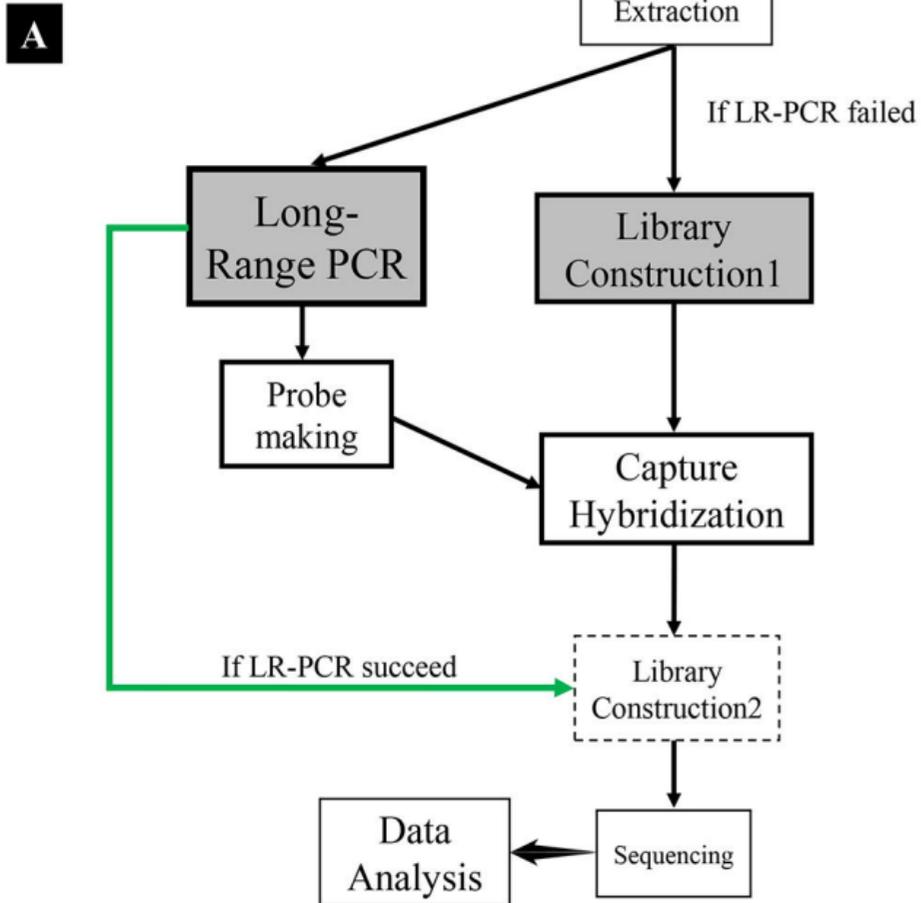
443 **Figure 1. A.** Schematic pipeline for enriching mitochondrial DNA to high-throughput sequencing. The  
 444 green line represents using the pair of LR-PCR amplicons to directly construct a library. Compared to  
 445 standard library, the LR-HY has modification in library construction one and two. 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.

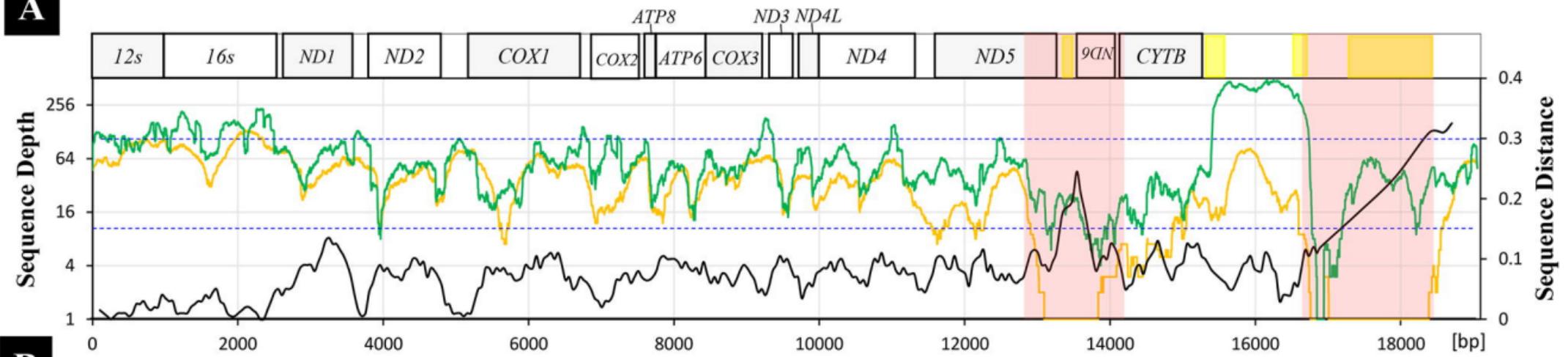
451 **Figure 2.** Coverage distributions for 400 bp and 2 kb library. **A** represents *Rana jiemuxiensis* results by  
 452 using standard capture hybridization (orange line) and LR-HY (green line). Black line represents DNA

453 sequence distance between *R. culaiensis* and *R. jiemuxiensis*. The sliding window length is 50 bp and the  
454 step length is 5 bp (below is the same). Dashed lines in A and B are constant at 0.15 and 0.3 sequence  
455 distance. The repetitive regions in *R. jiemuxiensis* which is labeled with yellow ranged from 13,424 to  
456 13,572 bp, 15,402 to 15,660 bp, 16,593 to 16,770 bp and 17,382 to 18,498 bp. **B** represents  
457 *Onychodactylus zhangyapingi* results by using standard capture hybridization (orange line) and LR-HY  
458 (green line). Black line represents DNA sequence distance between *R. culaiensis* and *O. zhangyapingi*.  
459 Dashed lines in A and B are constant at 0.15 and 0.3 of K2P. The regions with greatest sequence depth  
460 improvement are highlighted with red box. The five regions, i, ii, iii, iv, and v, with different sequence  
461 variation in *16s rRNA* are highlighted with blue box.

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

466



**A****B**