Isolation and characterization of DNA barcodes from distinctive and rare terrestrial animals in China using universal COI and 16S primers

Mali GUO, Zhenzhen PENG, Xiaoting ZHANG, Chaohai YUAN, Hanxi LU, Keyun ZHANG, Yafei CAI, Wei ZHANG

ABSTRACT

Background: Accurate taxonomic identification is the cornerstone for monitoring, conservation and management of ecological resources. China has the highest biodiversities and the richest species assemblages in the world, but is lacking in sufficient assessment to the abundant genetic variability. DNA barcoding is a proven tool employing sequence information for rapid and unambiguous species delineation. However, the ability of barcodes to distinguish species that are archaic and distinctive evolutionary lines remains largely untested.

Methods: In order to investigate the resources of terrestrial animals in China, regions from mitochondrial COI and 16S are barcoded for 395 specimens belonging to 54 selected species, many of which are indigenous representatives in danger. High success rate of PCR amplification is achieved by using universal COI and 16S primers with many numts pseudogenes co-amplified from mammalian samples.

Results: Application of barcodes to flag species is generally straightforward since no COI or 16S haplotypes are shared between closely related species. Barcoding gap, species resolution and phylogenetic relationships relying on our barcode libraries are further compared using distance and tree based approaches.

Conclusion: Results show that the discriminatory power of the two barcode markers could differentiate on a case-by-case basis, and also suggest a careful consideration of the nuclear numts for barcoding studies as they might provide a new understanding for evolution.

Keywords: DNA barcode · COI · 16S · Indigenous animals · China

Introduction

China is among the highest biodiversities and has the richest species assemblages in the world. It is estimated that over 10% of the world’s ecosystem types exist in this country, including ~2485 species of terrestrial vertebrates and at least 51,000 species of insects identified already (1). This species richness definitely is still underestimated as the rate of new and cryptic species discovery remains high. However, China’s genetic resources have also decreased sharply in the past decades due to its large human population and intensive human activities (2). Nearly half of China’s animals are found nowhere else and many are archaic and distinctive evolutionary lines nowadays at serious risk of extinction, such as the Giant panda (Ailuropoda melanoleuca) and Chinese alligator (Alligator sinensis) (3). Consequently, the high rates of
species discovery and loss have led to the urgent need for standardized methods to assess varied animal groups. Conventional taxonomic approaches for classification of species or breeds mainly rely on the characterization of morphological features, which are time-consuming and in some cases even lead to the disappearance of a species before its description (4). Moreover, both genetic and environmental factors underlie morphological variations. How genetic and environmental factors influence morphological characteristics remains a fundamental question under biological investigations (5). Observations based on morphology thereby are often unclear and challenging for species discovery and delimitation. Although whole genome sequencing and mass spectrometry-based protein profiling have also emerged as high throughput techniques allowing more precise species and breeds assessment (6, 7), the tedious analysis and high price slow their development in the field and, as a result, faster and easier alternatives with low costs are preferred.

DNA barcoding is a molecular tool employing sequence divergence in short and standardized gene regions to aid identification and discovery of species (8). It seeks to adopt one or few DNA fragments to efficiently and effectively assign any biological sample to its species regardless of the visual identification of the sample (9). The core idea is based on the fact that certain pieces of DNA, when aligned, can be found to vary merely to a limited degree within species while this variation is much less than between species (10). Therefore, whether samples of diverse species can be differentiated largely depends on the choice of the DNA sequence, which should be easy for amplification using universal PCR primers. Regions from mitochondrial genes usually form barcodes for members of animal kingdom. This is because each mitochondrion possesses insufficient DNA repair mechanisms and multiple copies of naked DNA without the protection of histone proteins, resulting in a 10-fold higher rate of nucleotide substitution in comparison to nuclear genome (11, 12). The accumulation of mutation in mitochondrial DNAs (mtDNAs) helps introduce more sequence diversity to establish phylogenetic relationships among animals and increases the chance to distinguish between closely related species (13).

A short fragment of ~648 base pairs (bp) at the 5’ end of the mitochondrial gene encoding the cytochrome c oxidase subunit 1 (COI) enzyme is the first and so far the most broadly used molecular marker for barcoding animals (14). It was reported that more than 95% of species in test assemblages of different animal groups, mainly insects, birds and fishes, showed characteristic COI sequences after successfully amplified using a universal pair of primers (15). More studies, however, challenged the degree of universality for COI and its primers for a number of reasons. For instance, the high variability of nucleotide sequences at the COI priming sites hinders its application to a broader spectrum of animal species (16). To address this issue, selected COI region and primers have been modified for barcoding species like amphibians (17). Yet how well the modifications work for bio-identification of other animals is still questionable, especially when coming to some distinctive and rare lines. As an alternative candidate, mitochondrial 16S ribosomal RNA gene is also often used, but its usage is substantially restricted to simple taxonomic analyses of microbiota (18). This is because of the prevalence of insertions and deletions in the non-coding RNA, which is thought to greatly complicate sequence alignments, although successes that 16S is superior to COI are also realized recently for Arthropoda and Amphibians (13, 19, 20). In spite of this, whether 16S could supply a sufficient resolution and robustness to discover entire animal kingdom has not been fully explored.

To date, little is known about the effectiveness of DNA barcode for evaluating taxonomic and phylogenetic structures of rare indigenous animals. In the paper present here, we select 54 representative species of distinctive terrestrial animals with 395 samples collected in 15 provinces throughout China, and systematically test the recovery of sequence information with universal primer sets that target short segments of the COI and 16S barcode regions. The goal of this work focuses on the prospect for investigating the genetic variability of threatened species using barcode sequences of COI and 16S through both distance-based and tree-based approaches. Our efforts will assist policy makers to understand the global patterns of biodiversity and persuade them to develop management strategies for prioritization and hot spots for conservation.

Materials and Methods

Sample acquisition
Animal samples including blood, semen, hair, tissues and faeces were collected in P. R. China following Animal Use Protocols approved by the Animal Care and Ethics Committee of Nanjing Agricultural University. Geographical distribution map of sample collection sites was created by making use of HyperText Markup Language 5 (HTML 5) and JavaScript scripting language (Figure 1). Blood samples were preserved with EDTA, whereas semen and faecal samples were frozen in liquid nitrogen. Hairs and tissues were kept in 75% ethanol. Species identity was based on morphological characters determined in the field. A total of 395 specimens representing 54 animal species, including 14 indigenous breeds of farm animals (4 species) in China, were used in this study (Supplemental Table S1). Some species were rare and represented by a single specimen solely, but for the majority multiple specimens were analyzed.

DNA extraction
DNA from blood samples was extracted using TINAamp Blood DNA Kit (DP348-03), while TINAamp Genomic DNA Kit (DP304-02) was adopted for DNA extraction from tissues. DNA from animal semen was purified using
Omega Forensic DNA Kit (D3591-02), and Omega Stool DNA Kit (D4015-01) was applied for camel faecal samples. All procedures were carried out according to the manufacturer’s protocol. DNA concentration and purity were assessed by Thermo Fisher Scientific NanoDrop One. DNA from animal hairs was extracted using alkaline lysis method as mentioned in (21). In brief, hair follicles from 10 hairs were boiled in 50 μL 0.2 M NaOH for 15 min, and 50μL Tris-HCl (pH = 6.0) was then added before the following experimental performance.

**COI and 16S amplification**

Less than 150 ng DNA or 5 μL hair lysate was used as template to amplify mitochondrial COI and 16S fragments. PCR was carried out in a 50 μL volume reaction using a Taq DNA polymerase (Vazyme, P213) with proofreading activity. Two sets of primers, COI-C0 and Chm4 (Table 1), were used depending on species to

---

**Table 1. Sequence information of universal PCR primers for COI and 16S.**

| Barcode | Primer | Name | Primer sequence 5’–3’ | Source |
|---------|--------|------|-----------------------|--------|
| **COI** | COI-C0 | COI-C02 | AYTCACAAATCATAAGATAGTTG | [1] |
| | | COI-C04 | AYCYYTGRGRTGACCAAAAAATCA | |
| | Chm4 | Chm4 | TYTCWACWAACAYAAAGAYATCGG | [1] |
| | Chm4 | Chm4 | AYCYYTGRGRTGRCRAARAAATCA | |
| **16S** | 16S | 16Sar-L | CGCCTGTTTATCAAAAAACAT | [2] |
| | 16Sbr-H | | CCGGCTGAACATCAGTCACG | |

[1] CHE, J., CHEN, H. M., YANG, J. X., JIN, J. Q., JIANG, K., YUAN, Z. Y., MURPHY, R. W. & ZHANG, Y. P. 2012. Universal COI primers for DNA barcoding amphibians. Mol Ecol Resour, 12, 247-58.

[2] PALUMBI, S. J. D. O. Z. & HAWAI, K. M. L. U. O. 1991. Simple fool's guide to PCR. Dept of Zoology & Kewalo Marine Laboratory University of Hawaii.
amplify the same barcoding region of COI (17). 16S
barcoding region was amplified with 16Sar-L and 16Sbr-H
primers (Table 1) by using 1.2 μM of each for all species
(22). PCR products were visualized on 1% (w/v) agarose
gels with ethidium bromide and recovered using TIANgel
Midi Purification Kit (DP209). Details of PCR conditions
and products for each sample are shown in Supplemental
Table S2.

TA cloning

Purified PCR products were linked to pMD19-T vector
with pMD™19-T Vector Cloning Kit (TaKaRa 6013)
in a 10μL reaction containing 4.75 μL PCR product,
0.25 μL vector and 5 μL Solution I. The mixture was
incubated at 16 °C for 30 min and then transformed into
DH5α competent cells (Vazyme, C502) according to the
instruction. After gentle mixing, competent cells were
placed on ice for 30 min, and then in 42 °C water bath
for 45s before quick transfer into ice for 3 min. Bacteria
were cultivated in 100 μL LB media at 37 °C for 10 min and
spread on LB agar plates containing 0.1mg/mL ampicillin,
0.04 mg/mL X-gal and 25 μg/mL IPTG for selection.
Single white colonies were picked up and cultivated for
sequencing.

Sanger sequencing

DNA barcoding regions were sequenced on both
directions using PCR primers directly for PCR products,
or using RV-M: GAGGGGATAAATTCACAGGG and M13F-47: CGCCAGGGTCTTCCGATCACGAC
primers for TA clonings. The number of colonies
sequenced is in line with one colony per 50 bp of
barcoding region unless identical barcode sequence was
sequenced twice. Sequencing results were assembled,
aligned and annotated using DNASTAR Lasergene
package before manually checked. All gained sequences
were confirmed by sequence similarity search available
in GenBank and deposited under the following accession
numbers: MZ046006-MZ046047, MZ046083-MZ046118,
MZ046726-MZ046731, MZ047098-MZ047179,
MZ048967-MZ049527, MZ050069-MZ050116,
MZ050118-MZ050213, MZ050493-MZ050515,
MZ061667-MZ061700, MZ068220-MZ068223,
MZ098871-MZ099441, MZ099449-MZ099455 for
COI sequences; MZ031856-MZ031915, MZ040165-
MZ040224, MZ040226-MZ040319, MZ040322-
MZ040403, MZ040406-MZ040487, MZ040500-
MZ040596, MZ040600-MZ040754, MZ040922-
MZ041011, MZ041035-MZ041093, MZ041115-
MZ041206, MZ042146-MZ042231, MZ042371-
MZ042467, MZ042537-MZ042621, MZ042714-
MZ042791, MZ042801 and MZ061594-MZ061630 for
16S sequences.

Sequence diversity

DNA Sequencing Polymorphism (DnaSP) was adopted
to screen for haplotypes and polymorphic sites by moving
a 200-bp-long sliding window 1 bp at a time, and to
calculate haplotype diversity (Hd), nucleotide diversity
(n) and the average number of nucleotide differences (K).
Parameters regarding the length and composition of DNA
and protein sequences were determined by MEGA
X. Translations of COI haplotypes were checked via
EditSeq with codon usage and isoelectric point calculated
automatically using genetic codes for mitochondrial DNA.

Distance and Tress matrices

DNA or protein sequences were aligned by the Clustal
W method before the following bioinformatic calculations
contained in the software package of MEGA X. Inter/
intraspecific genetic distance for COI and 16S was
plotted using the R programming language based on
Kimura’s 2-parameter (K2P) model. K2P is one of
the optimal theories for small pairwise distances, and is
widely used to calculate sequence divergences in DNA
barcoding literatures despite it is criticized (23). For each
animal class, the minimum interspecific distance was
compared with the maximum intraspecific distance in
order to determine the presence or absence of a barcoding
gap. Neighbor joining and maximum likelihood trees
construction were done using the “pairwise deletion”
option for the treatment of gaps and missing data. Node
support was computed by 1000 bootstrap replicates with
K2P distance for DNA data and Jones-Taylor-Thornton
(JTT) algorithms for proteins as a model of substitution.
Phylogenetic dendrograms were finally exported into the
web based iTOL (https://itol.embl.de/itol.cgi) tool for
annotation and visualization.

Statistical analysis

All data are expressed as the average ± SD. Spearman’s
correlation coefficient was used on data processing
to assess the relationship between the variables. Data
analysis was performed using the IBM SPSS software.
Comparisons between two groups were made using
Student’s t-test. The level of significance was set at p <
0.05.

Results and Discussion

Iso1ation and characterization of COI DNA barcodes

Both COI-C0 and Chm4 primer sets were reported to
span the same positions in the COI gene of amphibians
(17). The Chm4 primer set was designed to have more
2-fold degenerate bases in comparison to COI-C0 (Table
1). Therefore, the Chm4 primer set is utilized in this
study only when the COI-C0 primers fail to amplify a
perfect PCR product. In this experimental setting, 1480
COI sequences recovered from 391 species (> 98.98%
success) yield 787 unique haplotypes ranging from 616
to 663 bp with an average of 652 bp, among which 592
haplotypes are defined from 1094 sequences generated by
COI-C0 for 28/6 species/breeds, while the rest are from
25/8 species/breeds through Chm4 primer pair (Figure
2a, Supplemental Table S2 COI 1st pair and Table S3).
Notably, haplotype sequences of COI are only shared
among various breeds of farm animals, but not at the level of species.

The average nucleotide composition of COI haplotypes is 27.32% A, 30.58% T, 25.65% C and 16.45% G (Figure 2b). The GC content of these sequences varies from 25.60% to 52.40%, and identifies insects having the lowest GC content, consistent with what was reported before (24). As shown in Figure 2c, the density of variable sites in 200 bp windows along the alignment of COI haplotypes does not fluctuate widely for each animal group. Even so, the third position of genetic code generally is the least conserved and the second codon position shows slightly smaller variation than the first position (Figure 2d), in line with the “wobble phenomenon” during protein translation (25). However, this is not the case for 16S, which is a non-coding RNA and has all positions varying at an identical rate (Figure 2d’).

**Isolation and characterization of 16S DNA barcodes**

The 16S primers manage to obtain 1255 sequences from 390 out of 395 collected samples comprising 53/14 species/breeds in this study (Supplemental Table S2).
After removal of identical sequences within any one species, the barcode library is reduced to 592 unique haplotype segments with an average of 550 bp and lengths ranged from 503 to 574 bp (Figure 2a' and Supplemental Table S4). Of these, 89.02% haplotypes are newly discovered in current study after aligned to the 16S sequences in GenBank database. Similar to COI barcodes, there is no haplotype sequence of 16S found to be shared between species either. The base compositions of 16S are always biased towards A and T, which together are present in a higher proportion than GC, while the latter varies from 20.70% to 49.38% (Figure 2b'). As usual, insects have the lowest GC content. Molecular diversity indices of both COI and 16S barcodes are given in Table 2 for comparison. Overall, 16S shows lower nucleotides diversities (π) as well as average number of nucleotide differences (K) for all animal groups when compared with COI. Also distinct from COI, 16S haplotypes seem less polymorphic in their initial and final portions for all animal groups except for farm animals, which have the largest amount of variable sites owing to having too many 16S haplotypes for test (Figure 2c').

Isolation and characterization of COI protein peptides
Translation of COI haplotypes is also examined as COI is a protein-coding gene with essential function for oxidative phosphorylation (26). Results reveal that out of 787 unique COI barcodes, only 462 consisting of 51/11 species/breeds are functional domains containing no stop codon (Supplemental Table S3). Of those nonsense mutant fragments, more than 99% are novel sequences, majority of which are derived from mammalian samples (wild mammals and farm animals) (Figure 3a and b). In this respect, the percentage of new COI haplotypes able to encode proteins is >13% lower (86.36%), close to that of 16S (89.02%) when aligned to the data in GenBank.

The 462 functional COI barcodes mentioned above in total isolate 318 unique peptides, for all of which Leucine is the amino acid used most frequently although obvious codon bias for Leucine is observed among animal groups (Supplemental Table S5 and Figure 3c). The average length of the 318 COI peptides is 219 amino acids, within a narrow range from 214 to 220 amino acids, correlated with the small window of their isoelectric points between 4.059 and 6.308 (Figure 3d and e). When compared to data available in Genbank, around 87.11% peptides are characterized to be novel COI sequences (Figure 3b). However, sharing of peptide sequence is detected between species of not closely related birds or mammals, indicating COI proteins, in contrast to COI nucleotides, possess higher conservatism in the course of species evolution (Supplemental Table S5).

Barcoding gap and Species resolution
The primary application of barcode markers is to discriminate species. Methodological approaches for species delineation using DNA barcodes are commonly dependent on genetic distance-based measures, which rely on the assumption of a remarkable separation between intraspecific variation and interspecific divergence in the selected marker, also referred to as the barcoding gap (27, 28). K2P analysis thus is conducted using COI haplotypes, coding COI and 16S haplotypes to calculate
genetic distance respectively. The two COI datasets give comparable outcomes among all animal groups except for insects, in which coding COI is moderately better, displaying a murky gap with the minimum interspecific distance higher than the maximum intraspecific distance (Figure 4a-f'). Additionally both translatable COI and 16S haplotypes exhibit barcoding gaps for insects, amphibian, reptiles and birds, even though COI protein peptides are shared between some birds (Figure 4 and Supplemental Table S5). Nevertheless, both fail to show a separate distribution for mammals within and between species or breeds (Figure 4e-f').

Sequence datasets then are further evaluated using BLAST searches, which typically employ distance-based algorithms for pairwise alignments to assess species resolution (6). Performance of BLAST using COI peptides no surprise deliver the lowest species resolution because of the highest degree of sequence conservation (Figure 5a and Supplemental Table S5). As for DNA barcodes, COI overall offers better species resolution than 16S (Figure 5a),

https://doi.org/10.37175/stemedicine.v2i7.95
especially for insects, amphibians, reptiles and wild mammals (Figure 5b). Although the capability to make species-level identifications diverges tremendously for both DNA markers, a statistically significant correlation ($\rho = 0.76$, $p < 0.01$) is noted between the average resolutions of COI and 16S for each species or breed of farm animals (Figure 5c, Supplemental Table S3 and S4). In addition, BLAST results give a general impression that translatable COI barcodes have higher levels of species discrimination than their non-coding counterparts (Figure 5a). However, this is not always true since for species or breeds such as Golden monkey ($Rhinopithecus roxellanae$), Tibetan cattle ($Bos grunniens$), Sanhe horse and Tibetan horse ($Equus caballus$), some non-functional counterparts can perform even better (Supplemental Table S3).

**Phylogenetic relationships**

Contrary to divergence-based barcoding gap, construction of evolutionary trees places emphasis on conservation areas which are designed or prioritized according to their phylogenetic diversity (29). The most popular approaches to reconstruct phylogenies are the neighbor-joining (NJ) and the maximum likelihood (ML) algorithm. Although both phylogram constructions could correctly discriminate most COI haplotypes, with exception discussed below, NJ turns out to act better in light that it allocates related species closer to each other as hinted by the line colors representing animal classes in Figure 6. Yet the topologies of ML tree search sometimes can be more similar to the traditional taxonomic classifications. For example, rather than directly rooted from the same node as Sika deer ($Cervus nippon$) (Figure 6a),

**Table 2. Genetic diversity of COI and 16S barcodes generated in the present study.**

| Species            | barcode | N  | S    | H   | $\pi$ (SD) | Hd (SD) | K          |
|--------------------|---------|----|------|-----|------------|---------|------------|
| Insects            | COI     | 138| 363  | 62  | 0.9240 (0.0120) | 0.19422 (0.00208) | 126.246    |
|                    | 16S     | 122| 247  | 51  | 0.9010 (0.0160) | 0.18603 (0.00507) | 90.412     |
| Amphibians         | COI     | 37 | 261  | 19  | 0.9350 (0.0220) | 0.16619 (0.01298) | 108.853    |
|                    | 16S     | 42 | 194  | 20  | 0.9360 (0.0180) | 0.15242 (0.00657) | 80.632     |
| Reptiles           | COI     | 78 | 308  | 35  | 0.9050 (0.0210) | 0.15734 (0.00861) | 103.528    |
|                    | 16S     | 69 | 213  | 30  | 0.8880 (0.0270) | 0.14312 (0.00722) | 70.987     |
| Birds              | COI     | 54 | 448  | 26  | 0.9590 (0.0100) | 0.18004 (0.02340) | 108.92     |
|                    | 16S     | 69 | 213  | 44  | 0.9740 (0.0090) | 0.09766 (0.00357) | 53.030     |
| Wild Mammals       | COI     | 270| 405  | 157 | 0.9847 (0.0028) | 0.20070 (0.00152) | 127.703    |
|                    | 16S     | 236| 297  | 121 | 0.9775 (0.0038) | 0.12394 (0.00193) | 62.465     |
| Farm animals       | COI     | 903| 474  | 488 | 0.9610 (0.0041) | 0.14981 (0.00229) | 86.592     |
|                    | 16S     | 717| 368  | 326 | 0.9666 (0.0027) | 0.05204 (0.00205) | 27.684     |

N=number of sequences; S=number of polymorphic sites; H=number of haplotypes; Hd (SD)=haplotype diversity (standard deviation); $\pi$ (SD)=nucleotide diversity (standard deviation); K=average number of nucleotide differences.

Figure 5. Species resolution provided by COI, 16S DNA barcodes and COI peptides (a) for different animal groups (b) as indicated. Species resolution of unique sequences is predicted using BLAST searches based on reference libraries in Genbank, where reference sequences might be deposited unequally for various species. (***) $p<0.001$; (**) $p<0.01$; (N.S.) not significant. (c) Correlation between the average resolutions of COI and 16S for each species or breed of farm animals. Each dot is plotted according to the average values of resolution from all COI or 16S haplotypes belonging to the same species or breed. Spearman’s correlation coefficient $\rho = 0.76$ ($p<0.01$).
COI coming from cattle gather together and form a parallel branch with sheep before grouping with the branch containing deer according to the ML inference (Figure 6b). Unexpectedly, the Tibetan sheep (Ovis aries) haplotypes here are sequences bearing stop codons and considered as outgroup from the other 24 COI barcodes of sheep, whereas one coding sequence from Tianzhu white yak (cattle) is still misplaced with deer (Figure 6b box), highlighting a stronger discriminatory power of nonsense mutants that is usually ignored as reviewed by earlier studies (30, 31). Exceptional case common for both tree-building methods is Giant salamander (Andrias davidianus), whose coding peptides also fail to cluster with other Amphibians in protein phylogenies based on the JTT models (Supplemental Figure S1), indicating a special position of the archaic species in the evolutionary history (32).

It is noticeable that the NJ approach is much more superior to ML when it comes to 16S haplotypes (Figure 7). The average bootstrap proportion of NJ somehow is a little
bit lower than ML, but the NJ algorithm identifies all animal classes accurately with each individual 16S nearly perfectly assigned into their right sites. However, some 16S sequences originated from primates as well as Equus (horse and donkey) are found to be mixed within their groups and not arranged according to their species. These failures are also detected in phylogenetic trees constructed by COI, implying they have a very close kinship (Figure 6).

Different from the universal bar code consisting of a series of vertical bars that are printed on commercial products, DNA barcodes are genetic markers seeking to use sequence message of any biological sample, regardless of morphological identification of the sample, to address questions relating to taxonomy, ecology and evolution (33). A considerable amount of studies have evaluated the performance of a variety of barcode markers with respect to both their ease of PCR amplification and their capacity to delineate species (34). By making use of previously reported primers, here we manage to amplify and isolate DNA barcodes from nearly all Chinese terrestrial animals sampled. Totally 787 COI and 592 16S unique haplotypes are characterized in the survey of 54/14 animal species/breeds. Sequences of selected COI and 16S region are shared inside species but not between species with a low GC content on average, which is a typical feature of mitochondrial genome. In general, COI markers are longer and more polymorphic than 16S as reviewed by parameters of nucleotides diversities (π) and average number of nucleotide differences (K) (Table 2). These results clearly demonstrate that both COI and 16S primer pairs possess high degree of universality, competent to capture targeting segments efficiently from distinctive and rare lines of insects, amphibians, reptiles, birds and mammals across China.

In terms of choices of segments for barcoding and species diagnosis, DNA sequences that evolve slowly, like ribosomal genes in nucleus particularly, often do not differ among closely related organisms. Conversely, sequences that evolve rapidly may overwrite the traces of ancient affinities, but regularly reveal divergence between closely related species. Overwhelming evidence has suggested that mitochondrial genome is more likely to supply suitable candidate regions for barcoding animals for its maternal inheritance, non-introns and rapid evolution (35, 36). In order to assess the performance of mtDNAs in our context, distance-based and tree-based approaches are utilized to analyze the data libraries of COI and 16S garnered in this study. Substantially, all methods tend to be congruent with regards to success or failure. The opportunities to distinguish species dependent on the two markers are high for insects, amphibians, reptiles and birds with recognizable barcoding gaps, though COI appears to provide better species resolution after sequence alignment via BLAST, which might be a bias brought in by the reference databases available at this moment in Genbank (28). Furthermore, it will be more challenging to differentiate indigenous mammals using COI and 16S because neither of them is capable to display a perfect species boundary according to our exploration. This might be a reason why until now fewer barcode records of COI and 16S are accessible for mammals despite their prevalence in amphibians and reptiles. Nevertheless, it should be noted that the phylogenetic pattern reconstructed with 16S through NJ algorithm seems more reliable than COI, providing a picture generally analogous to the conventional taxonomic classifications. In sum, we speculate that both mitochondrial COI and 16S could function comparably at least as an eligible barcode marker for most species involved in current study, and that application of the two genetic markers relying on bioinformatic approaches should be cautious on a case-by-case basis.

Figure 8. Heteroplastic conditions in mitochondrial COI and 16S markers. COI and 16S are sequenced directly using PCR primers or after TA cloning, from which bacterial colonies are picked up in line with one colony per 50bp DNA unless the same sequence is captured twice. Heteroplasmy then is monitored either by the chance to obtain identical sequence (“heteroplasmic index” in a) or by the copy number of haplotypes generated (b). (b) Correlation between the copy numbers of COI and 16S barcodes for each individual specimen. Spearman’s correlation coefficient rho = 0.18 (p < 0.01).

https://doi.org/10.37175/stemedicine.v2i7.95
Table 3. COI sequences and haplotypes obtained from the same samples using two pairs of PCR primers.

| COI PCR | Sample ID | Jungle_fowl-B | Tibetan_donkey3-B | Tibetan_donkey5-B | Yunnan_donkey8-B | Yunnan_donkey9-B | Dezhou_donkey4-B | Dezhou_donkey52-S |
|---------|-----------|---------------|------------------|------------------|------------------|------------------|------------------|------------------|
| 1st pair | sequence # | 2 | 3 | 6 | 2 | 9 | 1 | 13 |
|         | haplotype # | 1 | 2 | 5 | 1 | 8 | 1 | 12 |
|         | translatable haplotype # | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
|         | sequence # | 4 | 4 | 2 | 3 | 3 | 6 | 8 |
| 2nd pair | haplotype # | 3 | 3 | 1 | 2 | 2 | 5 | 7 |
|         | translatable haplotype # | 3 | 2 | 1 | 1 | 2 | 3 | 3 |

On the other hand, in order to evaluate the potential misidentifications that could be caused by numts sequences, we decide to keep them in the analyses as outlined above. Our results show that coding COI in most cases allows methodological approaches to achieve more, but the influence of numts on the accuracy of taxonomic description is limited, suggesting that majority numts screened in this study are merely evolved for brief periods of time. Surprisingly, phylogenetic hierarchies deduced from both NJ and ML could position all COI numts to their expected places at the species level except for primates and Equus, which have been discussed earlier. More intriguingly, in-depth analysis in Figure 6 classifies two groups of non-coding numts from Tibetan sheep, among which seven are clustered with functional COI haplotypes whereas the other three are separated as outgroup with a stronger discriminatory power, at least in ML phylogram, than their coding counterparts. Taken together, these observations indicate that it is indeed not easy to characterize the role of numts as it stands for an ongoing evolutionary procession (41). Of course, it certainly is unfair to treat sequences as numts solely based on translation, which is not feasible for 16S. Yet it should be acceptable considering that the percentages of novel sequences are parallel among translatable COI, 16S and COI peptides but not non-coding COI (Figure 3b). In a word, our preliminary findings recommend that a careful investigation of numts may provide novel insights into the DNA barcoding system with potentially widespread scientific and practical benefits.

Acknowledgement
We sincerely appreciate Dr. Yanfen Cheng and Wenjia Huang at Nanjing Agricultural University for the sample collection. This work is funded by the National Key R&D Program of China (2018YFC1200201).

Conflict of interest
The authors declare that there is no conflict of interests regarding the publication of this paper.

Supplementary Information
The supplemental material can be downloaded online at: https://stemedicine.org/index.php/stem/article/view/95
