

# Evaluation and application of mitochondrial CO I gene in identification of endangered wildlife in multi-species mixed biological samples

CHEN Yun-xia<sup>1,2,3</sup>, XUE Xiao-ming<sup>1,2,3</sup>, HUANG Ya-lin, ZHOU Yong-wu<sup>1,2,3</sup>, HOU Sen-lin<sup>1,2,3</sup>, GUO Hai-tao<sup>1,2</sup>, JIANG Jing<sup>1,2</sup>

<sup>1</sup>Nanjing Forest Police College, Nanjing 210023;

<sup>2</sup>Forest Police Forensic Center of State Forestry Administration, Nanjing 210023;

<sup>3</sup>Key Laboratory of State Forest and Grassland Administration on Wildlife Evidence Technology, Nanjing 210023

**Abstract.** In this study, the second-generation high-throughput sequencing and DNA barcoding were combined to manually prepare multi-species mixed samples, and the mitochondrial gene CO I was used as a barcode to simultaneously identify the animal species in the mixed samples and identify endangered species. The results showed that under the family and genus level, the simultaneous detection rate of the species in the mixed samples was as high as 100%, and the species identification rate was as high as 89% at the species level, and with high sensitivity, as little as 1% of the trace species could be detected. However, nearly 30% of non-target classification annotations appeared at the species level. It can be concluded that the mini CO I barcoding can be applied to the simultaneous identification of animal species in mixed biological samples, and the species identification rate is high. Non-target classification match existing at the species level can be further improved by increasing the length of the barcoding, improving the sequencing technology, reference database and so on. In this study, DNA metabarcoding technology was used to evaluate the feasibility of identification of endangered animals in multi-species mixed biological samples with CO I, in order to lay a preliminary foundation for the advancement of DNA metabarcoding method in the field of wildlife forensic identification.

## 1 Introduction

Species genetic identification plays a key role in the investigation of illegal trade in endangered wildlife and food adulteration. Currently, DNA barcoding technology is an established molecular technique for species identification using standardized short DNA sequence. In recent years, DNA barcoding technology has been widely used in food supervision due to food labeling errors, food adulteration and food contamination<sup>[1-3]</sup>, such as traceability of processed foods such as seafood and meat products. The correct identification of the species contained in the food is essential to protect consumers from potential food adulteration, mislabeling of ingredients or food poisoning.

Another mature application of DNA barcoding in forensic science is to investigate illegal criminal activities such as illegal acquisitions, transportation, trading and smuggling of endangered wildlife. Species of flora and fauna are listed as endangered species by the Convention on International Trade in Endangered Species of Wild Fauna and Flora (CITES). Appendix I, II, and III list protected species based on the extent to which a population is threatened with extinction. In addition to regulated legal trade, a large part of the trade in endangered animals and plants is illegal. According to

statistics, from 2004 to 2013, the national customs criminal case investigated and dealt with 930 cases of smuggling of precious species, and seized 710 tons of precious species; nearly 10,000 illegal trade cases involving administrative investigations were investigated. In some cases, it is not difficult to identify the species of plants and animals involved, as long as the form is complete, it can be identified by traditional morphological classification. Identification is more difficult when only some animals or plants have no apparent morphological characteristics, or when plants or animals have been pulverized into articles. For samples that are unrecognizable in this form and do not know the composition of their species, a standardized, fast, and reliable method of identification is helpful and necessary for law enforcement. It is these advantages that make DNA barcoding the preferred method of identifying endangered species in animal and plant products.

In the case of species identification of traditional Chinese medicine, a mixed sample containing more than one component, DNA barcoding technology based on the generation of Sanger sequencing is not applicable. These samples usually contain multiple species, and these species can only be efficiently analyzed if multiple DNA barcoding can be sequenced in parallel, and second-generation sequencing technology can effectively

sequence these species. Therefore, this study combined the second generation of high-throughput sequencing with DNA barcoding technology (metabarcoding: DNA metabarcoding), and manually prepared mixed samples of three different species and different proportions, and used the mitochondrial gene CO I as a barcode to initially explore the feasibility and application prospect of DNA metabarcoding technology in rapidly identification of endangered species in complex samples.

## 2 Materials and methods

### 2.1 Materials

In this study, the animal samples used in the preparation of artificial mixed samples involved 11 genus of 9 families, including China's Cand secondary key protected animals, beneficial species with economic and scientific research values and animals supervised by CITES appendix. See Table 1 for specific information. All samples were taxonomically confirmed by animal morphology experts of Forest Police Forensic Center of State Forestry Administration.

**Table 1.** The source of animal samples

Species name	Protection level	Sample source
<i>Selenarctos thibetanus</i>	Second level, CITES Appendix I	Jingzhou Forest Public Security Bureau
<i>Manis pentadactyla</i>	Second level, CITES Appendix I	Wenzhou Customs Anti-smuggling Bureau
<i>Phasianus colchicus</i>	Beneficial species with economic and scientific research values	Nanjing Municipal Public Security Bureau Jiangning Branch
<i>Gloydus brevicaudus</i>		Nanjing Municipal Public Security Bureau Jiangning Branch
<i>Callosciurus erythraeus</i>	Beneficial species with economic and scientific research values	Changjiang Li Autonomous County Forest Public Security Bureau
<i>Cervus nippon</i>	First level	Zhangzhou Forest Public Security Bureau
<i>Muntiacus reevesi</i>	Beneficial species with economic and scientific research values	Shennongjia Nature Reserve Forest Public Security Bureau
<i>Naemorhedus goral</i>	Second level, CITES Appendix I	Shennongjia Nature Reserve Forest Public Security Bureau
<i>Mustela sibirica</i>	Beneficial species with economic and scientific research values	Rudong County Public Security Bureau
<i>Saiga tatarica</i>	First level, CITES Appendix II	Huai'an Public Security Bureau
<i>Lepus sinensis</i>	Beneficial species with economic and scientific research values	Nanjing Municipal Public Security Bureau Zweigstelle Jiangning

### 2.2 Methods

#### 2.2.1 Preparation of artificial mixed samples

**Table 2.** Species composition and percentage in the mixed samples

Sample number	Proportion of species in mixed samples										
	<i>Phasianus colchicus</i>	<i>Gloydus brevicaudus</i>	<i>Callosciurus erythraeus</i>	<i>Cervus nippon</i>	<i>Muntiacus reevesi</i>	<i>Naemorhedus goral</i>	<i>Mustela sibirica</i>	<i>Lepus sinensis</i>	<i>Selenarctos thibetanus</i>	<i>Saiga tatarica</i>	<i>Manis pentadactyla</i>
U1-1	11%	11%	11%	11%	11%	11%	11%	12%	11%	-	-
1-1G	9%	9%	9%	9%	9%	9%	9%	10%	9%	9%	9%
1-1W	11%	11%	11%	11%	11%	11%	11%	11%	1%	-	-

In this study, a total of three (U1-1, 1-1G, 1-1W) mixed samples were designed and prepared, each containing 9 to 11 species, and the components were mixed based on dry weight ratio (the relative concentration was 1% to 12%). Among them, the *Saiga tatarica* and the pangolin, which are often used for medicine, take their horns and nails as experimental materials, and these two hard materials are directly ground with a grinder; The remaining materials were taken from their muscle tissues, freeze-dried for 78 hours, and the lyophilized components were ground using an autoclaved mortar and pestle or a freeze mill, and then stored at -20°C. The individual components of the three mixed samples were weighed and thoroughly mixed by a tumbler for 20 hours and stored at -20°C until use. The specific species composition and proportion are shown in Table 2.

### 2.2.2 Genomic DNA extraction

In this study, the genomic DNA was extracted using the TaKaRa MiniBEST Universal Genomic DNA Extraction Kit Ver.5.0. During the extraction, the sample 1-1G contained bones components such as horns and nails, so the sample were digested overnight with metal bath, and the other samples were digested for 3 to 4 hours. The concentration and quality of DNA were determined by ultraviolet spectrophotometry and agarose gel electrophoresis.

### 2.2.3 Primer design

Due to the limitation of the sequencing length of the second-generation sequencing platform, and the consideration of foods, pharmaceuticals, etc., which are highly processed products, in which genomic DNA is highly degraded, mini barcoding are more suitable. Therefore, this study used a 360bp CO I fragment as a barcode [4], the primer sequence is as follows: CO Imini-1: 5'-GGWACGGWTGAACWGTWTAYCCYCC -3', CO Imini-2: 5'-TAIACYTCIGGRTGICCRAARAAYCA -3'.

### 2.2.4 Sequencing and analysis

The high-throughput sequencing of this study was completed by Jinweizhi Biotech's illumina HiSeq X Ten sequencing platform. The original data was analyzed by the software Bcl2fastq (v2.17.1.14) for image base recognition and preliminary quality analysis; The sequencing raw data was optimized using Cutadapt (v1.9.1), Qiime (1.9.1) and Vsearch (1.9.6) software; All sequences were subjected to OTU partitioning and statistical analysis of biological information using Qiime (1.9.1) and Vsearch (1.9.6) software.

## 3 Results and analysis

### 3.1 Statistical analysis of sequencing data

#### 3.1.1 Raw data statistical analysis

Data volume and sequencing quality for three sample sequencing raw data were counted. The number of reads of the data output of the 1-1G, 1-1W and U1-1 samples were 309838, 262030 and 270238, respectively. The specific data are shown in Table 3.

**Table 3.** Sequencing raw data information

Sample	Reads	Bases (bp)	Q20 (%)	Q30 (%)	GC (%)
1-1G	309838	77459500	94.41	91.93	43.81
1-1W	262030	65507500	94.51	92.08	43.71
U1-1	270238	67559500	94.59	92.15	43.63

Note: Sample: sample name; Reads: number of sequencing reads; Bases (bp): total number of bases; Q20 (%), Q30 (%): bases with Phred values greater than 20, 30, respectively, percentage of total bases; GC (%): The sum of the number of bases G and C as a percentage of the total number of bases.

#### 3.1.2 Sequencing data quality optimization

In high-throughput sequencing, sequencing errors such as point mutations usually occur, and the quality of the end of the sequence is relatively low. In order to obtain higher quality and more accurate bioinformatics analysis results, it's necessary to perform optimization processing such as removing primers and linker sequences and removing chimeras from the sequencing raw data to obtain the final effective data. The valid data statistics of the optimized and filtered samples of the three sample sequencing data are shown in Table 4.

**Table 4.** Filtered data quality statistics

Sample	PE reads	Nochimera	Avglen (bp)	GC (%)
1-1G	154919	150319	364.58	44.57
1-1W	131015	127335	364.60	44.49
U1-1	135119	131185	364.62	44.40

Note: Sample: sample name; PE\_reads: number of original PE reads; Nochimera: number of effective sequences after removal of chimera; AvgLen (bp): average length of effective sequence; GC (%): GC percentage content of valid data.

### 3.2 OTU analysis and species annotation

OTU is a unified mark that is artificially set for a certain taxonomic unit (genus, species, etc.) in the study of population genetics. In bioinformatics analysis, each sequence is from one species. To understand the species, genus and other information in a sample sequencing result, you need to classify the sequences. Through the categorization operation, the sequences were classified into many groups according to their similarity, and one group is an OTU. In this study, all sequences were subjected to OTU partitioning and statistical analysis of biological information at 97% similarity level. OUT cluster analysis was performed on three samples, resulting in a total of 48 OTUs. In order to obtain the species classification information corresponding to the

OTU, a representative sequence is selected for each OTU, and the representative sequence is subjected to species classification annotation using the RDP classifier, thereby obtaining the species composition of each sample. Among the three samples of 1-1W, 1-1G, and U1-1, 18.68%, 21.38% and 20.69% of the reads did not obtain species annotation information. The species classification of the three samples at the family level is shown in Table 5. The species classification at the genus level is shown in Table 6. The results show that the species identification rates of the samples 1-1W and U1-1 are 100% at the family and genus level. The species identification rate of sample 1-1G is 89%; however, the classification of two non-target species of the genus *Axis* and *Elaphurus* appeared at the genus level, which indicating that CO I is not high in species resolution of deer at the level under genus. The species classification of the three samples at the species level is shown in Figure 1. The results show that the species identification rates of samples 1-1W and U1-1 are both 89%, and the species identification rate of sample 1-1G is 73%. At the species level, in addition to the non-target species notes of the deer, *Mustela sibirica* was also incorrectly annotated as *Mustela erminea*. The results of this study also showed that *Selenarctos thibetanus* with a content as low as 1% in the sample 1-1W were also detected, but neither of the bone components in the sample 1-1G were detected.

**Table 5.** Taxonomy at the family levels

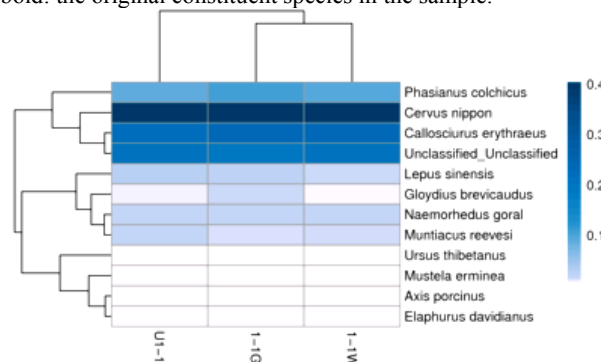
Taxon	1-1G	1-1W	U1-1
Cervidae	50036	49715	51577
Sciuridae	29032	29276	26516
Unclassified	22653	25929	25096
Phasianidae	12862	11605	10680
Leporidae	2604	1702	3847
Bovidae	2050	1962	1812
Viperidae	1777	915	1131
Mustelidae	176	166	422
Ursidae	102	22	211

Note: Taxon: the classification name of each species under the genus level; sample name: the number of reads of the sample in different species classification; the genus name displayed in bold: the original constituent species in the sample.

**Table 6.** Taxonomy at the genus level

Taxon	1-1G	1-1W	U1-1
<b>Cervus</b>	47759	47407	49001
<b>Callosciurus</b>	29032	29276	26516
Unclassified	22653	25929	25096
<b>Phasianus</b>	12862	11605	10680
<b>Lepus</b>	2604	1702	3847
<b>Naemorhedus</b>	2050	1962	1812
<b>Muntiacus</b>	1592	1657	1864
<b>Gloydus</b>	1777	915	1131
<i>Axis</i>	370	352	368
<i>Elaphurus</i>	315	299	344
<b>Mustela</b>	176	166	422
<b>Ursus</b>	102	22	211

Note: Taxon: the classification name of each species under the genus level; sample name: the number of reads of the sample in different species classification; the genus name displayed in bold: the original constituent species in the sample.



**Fig 1.** The heatmap of species cluster

Note: The column name is sample information, the row name is the species name, the tree above the graph is the sample clustering tree, and the left side of the graph is the species clustering tree. The value of each square with a different color in the middle heat map corresponds to the relative abundance value of each row.

## 4 Ddiscuses

When the DNA metabarcoding was first proposed in 2011, it was pointed out that its application value in the analysis of biodiversity. At present, this technology has been widely applied to all fields of metagenomics research, including soil<sup>[5]</sup>, ocean<sup>[6]</sup>, environmental pollution<sup>[7]</sup>, human gastrointestinal tract<sup>[8]</sup> and other directions.

In recent years, domestic and foreign scholars have also applied the DNA metabarcoding technology to the identification of mixed Chinese herbal ingredients<sup>[9]</sup>. In 2012, Coghlan *et al.* demonstrated the ability of DNA metabarcoding to detect species in complex Chinese medicine samples presented in the form of powders, crystals, capsules, tablets and herbal teas. Their findings showed that some Chinese medicine samples contain CITES-listed regulatory species, including Asian *Selenarctos thibetanus* and *Saiga tatarica*, as well as unlisted ingredients and potentially toxic and sensitizing plant constituents<sup>[10]</sup>. In 2014, Cheng *et al.* based on the Liuwei Dihuang Pill formula widely used in China, carried out macro bar code analysis on traditional Chinese medicine preparations with clear composition. The results show that there are significant differences in quality and safety between different commercial Chinese medicine preparations, because the unlisted *Acanthopanax senticosus* in some preparations may pose a safety risk to consumers<sup>[11]</sup>. In 2017, Raclariu *et al.* used DNA metabarcoding technology combined with HPLC-MS to conduct *Veronica officinalis* and its commonly used adulterants *V. chamaedrys* in 16 medicinal products of *Veronica officinalis* identification and identification research and found that only 15% of the products were detectable by *Veronica officinalis*, and 62% of the products detected mixed *V. chamaedrys*<sup>[12]</sup>. In 2018, Li *et al.* used the Ion S5 high-throughput

sequencing platform to identify and analyze the species of the commercially available Ruyi Golden Powder containing 10 herbs, and detected the ITS2 sequence of 8 herbs<sup>[13]</sup>. These studies show that the DNA metabarcoding method is a more effective method for reviewing highly processed Chinese medicine products and will help to monitor its legitimacy and safety.

In this study, three mixed samples of multi-species were prepared manually. U1-1 was a mixed sample with a mass ratio of 1:1 for all components, 1-1W was a mixed sample containing trace components (1%), and 1-1G was a mixed sample containing high nose antelope horn and pangolin powder. High-throughput sequencing technology was used to evaluate and analyze the feasibility of mitochondrial CO I as a barcode in identifying endangered animals in complex biological samples. The results showed that at the level of family, genus and species, the simultaneous identification rates of the species in the mixed samples were as high as 100%, 100%, 89%, respectively, and the sensitivity was high. The specie with the content as low as 1% was detected at the same time. However, bone components with low DNA content such as animal horns and scales were not detected. In this study, although DNA metabarcoding technology based on CO I can achieve simultaneous recognition of animal species in mixed samples, and the detection rate is as high as 89%, at the species level, nearly 30% of non-target classification matches are also produced. This may be due to the fact that the mini CO I barcode is used in this study. It contains less classified information and low resolution. At the same time, the misreading of some bases introduced during the sequencing process will also lead to the classification of false positives. However, with the further development of the second-generation sequencing technology, the increase in sequencing length and the improvement of sequencing accuracy, such problems will be greatly improved.

Based on this study, DNA metabarcoding has great potential in the field of identification of Chinese medicines and identification of food-derived components. However, there are still many difficulties to overcome, such as highly processed foods, high quality DNA extraction in pharmaceuticals, low resolution of mini barcoding, and the quality and integrity of reference sequence databases, which are especially important for law enforcement issues. Therefore, the application of DNA metabarcoding to the identification of forensic wildlife samples is still in its infancy, and it faces many challenges. This study hopes to lay a preliminary foundation for the advancement of DNA metabarcoding method in the field of animal and plant forensic identification.

## Acknowledgment

This work is financially supported by the Fundamental Research Funds for Central Universities (LGZD201908).

## Author

Chen Yunxia (1982-), female, Han nationality, doctor, associate professor, mainly engaged in the identification and protection of endangered wild animals and plants, E-mail: yunaini@163.com

## References

1. Fajardo V, González I, Rojas M, García T, Martín R. A review of current PCR-based methodologies for the authentication of meats from game animal species[J]. Trends Food Sci Technol. 2010;21(8): 408–21.
2. Wong EH-K, Hanner RH. DNA barcoding detects market substitution in North American seafood[J]. Food Res Int. 2008;41(8):828–37.
3. Hanner R, Becker S, Ivanova NV, Steinke D. FISH-BOL and seafood identification: geographically dispersed case studies reveal systemic market substitution across Canada[J]. Mitochondrial DNA. 2011;22(sup1):106–22.
4. Staats M, Arulandhu A J, Gravendeel B, et al. Advances in DNA metabarcoding for food and wildlife forensic species identification[J]. Analytical and Bioanalytical Chemistry, 2016, 408(17):4615-4630.
5. Gillespie D E, Brady S F, Bettermann A D, et al. Isolation of antibiotics turbomycin A and B from a metagenomic library of soil microbial DNA[J]. Appl Environ Microb,2002,68(9):4301.
6. [6] Venter J C, Remington K, Heidelberg J F, et al. Environmental genome shotgun sequencing of the Sargasso Sea[J]. Science,2004,304(5667):66.
7. Wagner M, Nielsen P H, Loy A, et al. Linking microbial community structure with function: fluorescence in situ hybridization-microautoradiography and isotope arrays[J]. Curr Opin Biotech,2006,17(1):83.
8. Gill S R, Pop M, DeBoy R T, et al. Metagenomic analysis of the human distal gut microbiome[J]. Science,2006,312(5778):1355.
9. Gao Y Z, Wei J, Liu Z W, Zhou J. Application of DNA metabarcoding technology in identification of Chinese patent medicines [J]. China Journal of Chinese Meteria Medica, 2019,44(2):261-264.
10. Coghlan M L, White N E, Murray D C, et al. Metabarcoding avian diets at airports: implications for birdstrike hazard management planning[J]. Investigative Genetics,2012,4(1):27.
11. Cheng X W, Su X Q, Chen X H, et al. Biological ingredient analysis of traditional Chinese medicine preparation based on high-throughput sequencing: the story for Liuwei Dihuang Wan[J]. Sci Rep,2014,4.
12. Raclariu A C, Mocan A, Popa M O, et al. Veronica officinalis product authentication using DNA metabarcoding and HPLC-MS reveals widespread

- adulteration with *Veronica chamaedrys*[J]. *Front Pharmacol*,2017,8:378.
13. Li Q, Sun Y, Guo H, et al. Quality control of the traditional Chinese medicine Ruyi jinhuang powder based on high-throughput sequencing and real-time PCR[J]. *Sci Rep*,2018,8(1): 8261.