Sequencing and Structural Characteristic Analysis of Mitochondrial Genome in Zhijin White Goose(Anser cygnoides)

Zhi-Li Guo
Key Laboratory of Animal Genetics, Breeding and Reproduction in the Plateau Mountainous Region, Ministry of Education, Guizhou University

Yong Zhang (gzuzy@139.com)
Key Laboratory of Animal Genetics, Breeding and Reproduction in the Plateau Mountainous Region, Ministry of Education, Guizhou University

Hong Yang
Bijie Animal Husbandry Station

Tian-Song Wang
Tongren Polytechnic College School of Agronomy

Wei-Wei Wang
Key Laboratory of Animal Genetics, Breeding and Reproduction in the Plateau Mountainous Region, Ministry of Education, Guizhou University

Shan-Shan Zeng
Key Laboratory of Animal Genetics, Breeding and Reproduction in the Plateau Mountainous Region, Ministry of Education, Guizhou University

Yong Guo
Guizhou Cattle Industry Group Science & Technology Services Co., Ltd

Li Ye
Bijie Animal Husbandry Station

Ao Du
Bijie Animal Husbandry Station

Zhi-Wei Wang
Key Laboratory of Animal Genetics, Breeding and Reproduction in the Plateau Mountainous Region, Ministry of Education, Guizhou University

Su-Mei Zeng
Key Laboratory of Animal Genetics, Breeding and Reproduction in the Plateau Mountainous Region, Ministry of Education, Guizhou University

Jian Yuan
Key Laboratory of Animal Genetics, Breeding and Reproduction in the Plateau Mountainous Region, Ministry of Education, Guizhou University

Lei Wang
Bijie Animal Husbandry Station

Method Article

Keywords: Zhijin White Goose, Mitochondrial Genome, Variety identification, Phylogeny, Structural Features

Posted Date: April 13th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1540661/v1

Abstract

Zhijin White Goose has the characteristics of fast growth, good fattening performance, fresh and high-quality meat, rough feeding resistance, and strong disease resistance. It is a precious white goose resource unique to Guizhou Province, China. However, the number of Zhijin White Geese is decreasing endangered state. The purpose of this study to establish a breed conservation standard of Zhijin White Goose by using mitochondrial genome(mtGenome), and preliminarily analyze the mtGenome. In this study, the mtGenome sequence (GenBank MZ427898) of Zhijin White Goose was obtained by the high-throughput sequencing method. The mtGenome differences between the Zhijin White Goose and other different geese breeds were analyzed, and determining the phylogenetic position of the Zhijin White Goose in Anseriformes. The structural characteristics of some sequences and the secondary structural characteristics of RNA were analyzed, which in order to provide some theoretical reference for the follow-up functional research. The results demonstrated that the length was 16 739 bp, and the base composition was A (30.22%), G (15.08%), C (32.23%), and T (22.47%). It contained 13 protein-coding genes (PCGs), 22 tRNAs, 2 rRNAs, and 1 D-loop region. The AT content of the mtGenome and PCGs in Zhijin White Goose was 51.58% and 52.69%, respectively, which showed AT preference. The 22 tRNAs, except tRNA-Ser $^{(AGY)}$, form a typical clover structure. Phylogenetic tree analysis based on the mtGenome revealed that the Zhijin White Goose is closely related to the Swan Goose and Chinese Goose and is distantly related to breeds outside China. In conclusion, the mtGenome can be used as a standard to distinguish Zhijin White Goose from other goose breeds, and it can be used as a breeds, candidate identification criterion for Zhijin White Goose.

Introduction

Eukaryotic cells carry two genomes: the nuclear and mitochondrial genomes (mtGenome). These two genomes are independent in terms of replication, isolation, and inheritance (Jayaprakash *et al.*, 2015). The animal mtGenome is usually a small (15–20 kb) double-stranded, maternal genetic ring. It is characterised by a simple structure, multiple copies, independent replication, high coding efficiency, fast evolution rate, and no tissue specificity. It plays an important role in metabolism, programmed cell death, disease, and ageing and contains important phylogenetic information (Lavrov and Pett, 2016). Compared with nuclear genes, the mtGenome is conserved in gene content, reaches high levels, and has no introns. Thus, it is widely used as a barcode to track the history of phylogenetic evolution (Cao *et al.*, 2006). Some mtGenome domains, such as the protein-coding genes *Cytb*, *ND2*, and *COI*, have special significance. As the speed of evolution was moderate, these genes, together with the control region, have been widely used to solve taxonomic problems of controversial biological groups, especially the phylogenetic relationships of some birds (Liu *et al.*, 2013, Slack *et al.*, 2007). Owing to the revolutionary progress in molecular technologies related to PCR, sequencing, and data analysis, the integrity of the mtGenome has attracted increasing attention (Pereira and Baker, 2006). Due to the increasing information on mtGenome sequence data used for phylogenetic relationship research at the species and population levels(Boore, 2004), the mtGenome has become a very useful molecular marker for reconstructing phylogenetic relationships of different animal classification levels. Complete mtGenome analysis can provide sequence information for phylogenetic research and further clarify its structure and function(Harrison *et al.*, 2004).

The Zhijin white goose is a rare local breed of white geese in Guizhou Province, China, and is an important poultry genetic resource. In recent years, the progress of breed conservation has been slow. In addition, hybridisation with newly introduced breeds in the production area, and factors such as strong nesting behaviour and weak fecundity, resulted in performance degradation. The production yield decreases yearly, and it is in an endangered state. The protection of Zhijin white goose resources is imminent.

Materials And Methods

Sample collection

One adult Zhijin white goose produced by the Ruixiang farm in Zhijin County, Guizhou Province, was selected. Blood was collected from the inferior wing vein for DNA extraction and sequencing.

Genomic DNA extraction

The Ezup Column Blood Genomic DNA Purification Kit (Shanghai Shenggong: Ezup Column Blood Genomic DNA Purification Kit) was used to extract genomic DNA following the manufacturer's instructions. DNA concentration was determined using a nanodrop, a micro nucleic acid analyser, to ensure that it could be used for subsequent library construction and sequencing.

Library construction and high-throughput sequencing

A paired-end sequencing library with an insert size of 350bp was constructed according to the Illumina DNA library construction standard process. After library construction, quality control was performed using qPCR and an Agilent 2100 Bioanalyzer (Agilent Technologies, USA). The DNA library was sequenced using the Illumina Novaseq6000 (Illumina, USA) high-throughput sequencing platform and a paired end 150 bp (PE150) sequencing strategy.

Sequence assembly and quality control

The mtGenome sequence was analysed using SPAdes v.3.5.0 software (http://cab.spbu.ru/software/spades/) to perform splicing and assembly. The geese reference genome was downloaded from NCBI. The mtGenome sequence was separated from the total DNA sequence using sequence alignment, and the extracted mtGenome sequence was subjected to statistical data analysis and quality control.

Sequence annotation and genome analysis

MitoFish (http://mitofish.aori.u-tokyo.ac.jp/) and ORF Finder (https://www.ncbi.nlm.nih.gov/orffinder/) were used to annotate the mtGenome(Iwasaki et al., 2013). Tandem Repeats Finder Database v.4.09 (Benson, 1999) was used to identify repeated domains. tRNA annotation was performed using ARWEN(Version1.2 http://mbio-serv2.mbioekol.lu.se/ARWEN/) (Laslett and Canback, 2008). The Mfold software (http://unafold.rna.albany.edu/?q=mfold/DNA-Folding-Form) (Zuker, 2003) was used to predict the stem-loop structure. The circular structure of the mtGenome was drawn using OGDRAW(Lowe and Chan, 2016), and the nucleotide composition was analysed using MEGA 6.0. The complete sequence of the mtGenome was annotated and saved in GenBank accession number MZ427898.

Analysis of structural characteristics of tRNA and rRNA

The Auto Traveler software on the RNAcentral website (https://rnacentral.orghelpsecondary-structure) to map the rRNA secondary structure(The RNAcentral Constortium, 2018). The tRNA scan-SE software was used (http://lowelab.Ucsc.edu/tRNAscan-SE/) to predict the position and secondary structure of tRNA (Lowe and Chan, 2016), together with the VARNA software to draw the secondary structure map(Darty et al., 2009). Relative synonymous codon usage (RSCU) was calculated as r to evaluate codon usage preference, the equation was as follows:

$$\mathrm{RSCU}_{ij} = \frac{x_{ij}}{\frac{1}{n}\sum_{j=1}^{n_i} x_{ij}}$$

Where xij = the occurrence times of the jth codon encoding the ith amino acid, n = the number of synonymous codons encoding the ith amino acid (values are 1 ~ 6). If there is no codon usage preference, the RSCU value is 1; if the codon is used more frequently than other codons, the RSCU value is greater than 1; otherwise, it is less than 1.

The chips software on the EMBOSS website (https://www.bioinformatics.nl/emboss-explorer/) was used to calculate the effective number of codons (ENC) of the protein-coding gene, and the cusp software was used to calculate the GC content of the third base of the codons(GC3s). The ENC-GC3s distribution (Sharp and Li, 1986) was plotted using GraphPad Prism 8 (version 8, GraphPad, La Jolla, CA, USA).

Phylogenetic tree construction

Thirteen geese mtGenome sequences were downloaded from GeneBank, including Sichuan White Goose (SW, MK133022), Hepu Goose (HP, KP943133), Youjiang Goose (YJ, KP881611), Wugang Tong Goose (White) (WG, KP026178), Swan Goose (SG, KJ124555), Landes Goose (LD, MK133021), Roman White Goose (RW, EU932689), Xupu Goose (XP, KJ94188), Zhedong White Goose (ZD, KT427463), Mayang White Goose (MY, MK102803), Bar-headed goose (BG, KM455570), Taiga bean Goose (TG, HQ890328), and Yanling Goose (YL, KJ778677). The MEGA 6.0 software Kimura two-parameter model(Nishimaki and Sato, 2019) was used to calculate the genetic distance between the 13 breeds and estimate their divergence time by building a phylogenetic tree using the Neighbor-joining (NJ) method and a Bootstrap value of 1,000. All other settings were the system defaults.

Results

Screening and Analysis of mtgenome Data of Zhijing White Goose

The total DNA of Zhijin white geese was analysed by high-throughput sequencing (Table 1). A total of 7.02 million original reads were obtained, and the base number of the original reads for double-ended sequencing was 1.05 billion. From the quality inspection, the Q20 value was 93.08–97.30%, and the Q30 value was 86.47–93.88%, and the data could be used for subsequent isolation of the mtGenome.

Table 1
Statistics of total DNA sequencing data of Zhijin White Goose

Sample		Length(bp)	Reads	Base(bp)	Total Base(bp)	Q20(%)	Q30(%)
Zhijin white goose	Read1	150	3,514,442	527,166,300	1,054,332,600	97.30	93.80
	Read2	150	3,514,442	527,166,300		93.88	86.47
Note:Read1: the first read of double ended sequencing;							
Read2: the second read of double ended sequencing;							
Length (bp): reading length of sequencing sequence;							
Reads: the number of reads in the original data;							
Base (bp): the number of sequenced bases in the original data;							
Total base (bp): the total base number of sequencing in the original data							
Q20: proportion of alkali base with mass value greater than 20 (error rate less than 1%) in total alkali base							
Q30: proportion of alkali base with mass value greater than 30 (error rate less than 0.1%) in total alkali base							

Table 1 here

The mtGenome sequences of the samples were separated, and the extracted mtGenome sequences were subjected to statistical data analysis and secondary quality detection (Table 2). Finally, 6006 clean readings were obtained, the number of bases was 0.9 million, the Q20 value was 93.82–98.05%, the Q30 value was 84.83–93.24%, and the guanine (G) and cytosine (C) bases accounted for 47.58% and 47.29% of the total bases, respectively. The average depth of mtGenome sequencing was 53x. When the isolated readings were aligned with the goose mitochondrial reference genome, the mitochondrial coverage was 100%. These results showed that the experimental sampling procedures and conditions met the sequencing requirements, and the results could be used for subsequent biological information analysis.

Table 2
Statistics of mtGenome sequencing data of Zhijin White Goose

	Length(bp)	Reads	Base(bp)	Total Base(bp)	Q20(%)	Q30(%)	GC(%)	Genome depth(X)	Genome coverage(%)
Read1	150	3,003	450,450	900,900	97.30	93.80	47.05	53	100
Read2	150	3,003	450,450		93.88	86.47	47.29		
Note:Read1: the first read of double ended sequencing;									
Read2: the second read of double ended sequencing;									
Length (bp): reading length of sequencing sequence;									
Reads:the number of sequenced reads;									
Base (bp): the number of sequenced bases;									
Total base (bp): the total base number of sequencing;									
Q20: proportion of alkali base with mass value greater than 20 (error rate less than 1%) in total alkali base;									
Q30: proportion of alkali base with mass value greater than 30 (error rate less than 0.1%) in total alkali base;									
GC%: the proportion of G base and C base in the total base;									
Genome depth (x): the average depth of mitochondrial genome sequencing.									
Genome coverage (%): mitochondrial coverage.									

Table 2 here

Annotation information of the whole mtGenome of Zhijin White Goose

The complete mtGenome length of Zhijin White Goose was 16 739 bp, and included 13 protein-coding genes (PCGs), 2 ribosomal RNA (rRNA) genes, 22 transfer RNA (tRNA) genes, and a major non-coding control domain (D-loop region), as shown in Fig. 1. Table 3 shows the length, start and end positions, coding chains, protein codes of all rRNA genes, tRNA genes, protein-coding genes, and non-coding domains of Zhijin White Goose Information, such as the start codon and stop codon of the gene, and their varying values. No tandem repeat series was found in the mtGenome sequence of Zhijin White Goose.

Table 3
The structural characteristics of the mtGenome of Zhijin White Goose

Gene name	starting point	End position	Length/bp	Coding chain	Start password	End password	Interval/Repetition	Interval sequence
D-loop	1	1175	1175	H			0	
tRNA-Phe	1176	1243	68	H			0	
12S-rRNA	1244	2231	988	H			0	
tRNA-Val	2232	2302	71	H			0	
16S-rRNA	2303	3912	1610	H			0	
tRNA-Leu$^{(UUR)}$	3913	3986	74	H			0	
NAD1	3993	4970	978	H	ATG	AGG	6	TCACCC
tRNA-Ile	4969	5041	73	H			-2	
tRNA-Gln	5049	5119	71	L			7	GTTGACC
tRNA-Met	5119	5187	69	H			-1	
NAD2	5188	6228	1041	H	ATG	TAG	0	
tRNA-Trp	6227	6299	73	H			-2	
tRNA-Ala	6305	6373	69	L			5	CACAT
tRNA-Asn	6375	6447	73	L			1	C
tRNA-Cys	6451	6516	66	L			3	CAA
tRNA-Tyr	6517	6586	70	L			0	
CO1	6588	8138	1551	H	GTC	AGG	1	C
tRNA-Ser$^{(UCN)}$	8130	8202	73	L			-9	
tRNA-Asp	8205	8273	69	H			2	AT
CO2	8275	8961	687	H	GTG	TAA	1	T
tRNA-Lys	8963	9031	69	H			1	T

Note: H (Heavystrand) stands for heavy chain, L (Lightstrand) stands for light chain.

Interval/Repetition: A positive number indicates that there is an interval sequence between adjacent genes, and a negative number indicates that there is overlap between adjacent genes.

Gene name	starting point	End position	Length/bp	Coding chain	Start password	End password	Interval/Repetition	Interval sequence
ATP8	9033	9200	168	H	ATG	TAA	1	T
ATP6	9191	9874	684	H	ATG	TAA	-10	
CO3	9874	10657	784	H	ATG	T-	-1	
tRNA-Gly	10658	10726	69	H			0	
NAD3	10727	11078	351	H	ATG	TAA	0	
tRNA-Arg	11080	11150	71	H			1	C
NAD4L	11151	11447	297	H	ATG	TAA	0	
NAD4	11441	12818	1378	H	ATG	T-	-7	
tRNA-His	12819	12887	69	H			0	
tRNA-Ser$^{(AGY)}$	12888	12954	67	H			0	
tRNA-Leu$^{(CUN)}$	12954	13024	71	H			-1	
NAD5	13025	14842	1818	H	GTG	AGA	0	
Cytb	14850	15992	1143	H	ATG	TAA	7	CTCACTA
tRNA-Thr	15995	16062	68	H			2	AT
tRNA-Pro	16071	16139	69	L			8	CCCCAAAC
NAD6	16150	16671	522	L	ATG	TAG	10	ACCCTAACCC
tRNA-Glu	16672	16739	68	L			0	
Note: H (Heavystrand) stands for heavy chain, L (Lightstrand) stands for light chain.								
Interval/Repetition: A positive number indicates that there is an interval sequence between adjacent genes, and a negative number indicates that there is overlap between adjacent genes.								

Figure 1 here

As shown in Table 3, the Zhijin White Goose mtGenome contained 15 domains with interval sequences, the longest is between tRNA-Pro and *NAD6* (10 bp). Seven domains had overlapping sequences, and the longest was 10 bp between *ATP8* and *ATP6*. The control area (1 393 bp) was flanked by tRNA-Glu and tRNA-Phe. The base composition was A (30.22%), G (15.08%), C (32.23%), and T (22.47%), and the A + T content (52.69%) was higher than G + C content (47.31%), indicating that the mtGenome of Zhijin White Goose has AT preference.

Table 3 here

Structural characteristics of rRNA and tRNA

The mitochondria of Zhijin white geese contain two rRNAs, 12S rRNA and 16S rRNA, located on the H chain. The position of 12S rRNA in the mitochondrial sequence was 1 244–2 231 bp, and the position of 16S rRNA in the mitochondrial sequence

was 2 303–3 912 bp. The tRNA-Val gene was sandwiched between the two rRNAs. The secondary structure of the mitochondrial 12S rRNA was drawn based on the *Cygnus melancoryphus* mitochondrial 12S rRNA (b. 16. m. C. melancoryphus) file provided by CRW(Cannone *et al*., 2002). In this secondary structure, domains I and II are variable regions, whereas domains III and IV are conserved (Fig. 2). The secondary structure of mitochondrial 16S rRNA was drawn with the human (*Homo sapiens Linnaeus*) mitochondrial 16S rRNA (mHS_LSU_3D) file as the template. This secondary structure is composed of six domains, of which the I–III and VI domains are variable regions, and the IV and V domains are conserved regions (Fig. 3).

Figure 2 here

Figure 3 here

On comparing the secondary structure diagrams of 12S rRNA and 16S rRNA, the 12S rRNA gene was found to be more conserved than the 16S rRNA gene. The secondary structure diagram also showed that the base sequence of the stem region of the rRNA is more conserved than the base sequence of the loop region, and there are fewer base substitutions, insertions, and rearrangements.

As shown in Table 3, there were 22 tRNA genes, ranging in length from 66–74 bp, and the full length of tRNA was 1,540 bp, accounting for 9.21% of the entire genome. There were two tRNA-Ser (UCN and AGY) and two tRNA-Leu (UUR and CUN). Fourteen tRNAs were located on the heavy chain (H-chain), and eight tRNAs were on the light chain (L-chain). The longest tRNA gene was tRNA-Leu$^{(UUR)}$ and the shortest was tRNA-Cys. The secondary structure prediction results showed that the 21 tRNAs of the mtGenome in Zhijin White Goose could form a clover-type secondary structure (Fig. 4), whereas tRNA-Ser$^{(AGY)}$ cannot form a clover structure due to the deletion of the dihydrouracil arm (D arm) (Fig. 4R). In addition, 22 mismatches were found, all were G = U pairings, and they appeared on each arm, 8 in the amino acid arm (AA arm), 5 in the dihydrouracil arm (D arm), 3 in the anticodon arm, and 6 on the pseudouracil arm (TψC arm).

Figure 4 here

D-loop region

The length of the D-loop region of the mitochondria of Zhijin White Goose was 1 175 bp. The G + C content was 46.13%, lower than the A + T content (53.87%), showing AT preference. Comparative analysis showed that, consistent with other vertebrates, the D-loop region can be divided into three domains: the terminal associated sequences (TAS), the central domain (CD), and the conserved sequence blocks (CSB)(Randi and Lucchini, 1998). According to the reported structural characteristics of the avian D-loop region (Quinn and Wilson, 1993), the relevant segments of the mitochondrial D-loop region of Zhijin White Goose were identified (Fig. 5). The length of the TAS region in the mitochondrial D-loop region of Zhijin White Goose is 362 nt, which contains the highly conserved 5'-GTGCAT-3' motif of the avian TAS region. A sequence highly similar to the “goose hairpin” sequence unique to geese was also identified. The secondary structure of this sequence was also composed of one stem consisting of seven C/G pairs and one loop containing the TCCC motif (Fig. 6). The CD region in the mitochondrial D-loop region of Zhijin White Goose spans 466 highly conserved nucleotides and has four key sequence segments: CSB-F, CSB-E, CSB-D, and CSB-C. The CSB region of the Zhijin white goose is 346 nt long. Only the conserved sequence cbs-1 was identified in the CSB region of Zhijin white geese, whereas CSB-2 and CSB-3 were not identified.

Figure 5 here

Figure 6 here

Protein-coding genes

The mitochondria of Zhijin White Goose contained 13 identical protein-coding genes. The length of the genes ranged from 168 bp to 1 818 bp and the total length of the genes reached 12 661 bp, accounting for 75.64% of the total length of the genome. *NAD6* was encoded by the L-chain while the remaining 12 PCGs were encoded by the H chain. As shown in Table 3,

among the 13 PCGs, 10 used ATG as the promoter, 2 used GTG as the promoter, and *CO1* used GTC as the promoter. Six PCGs used TAA as the stop codon, two PCGs (*NAD2* and *NAD6*) used TAG as the terminator, two PCGs used AGG as the terminator, *NAD5* used AGA as the terminator, and *CO3* and *NAD4* had incomplete stop codons (T-).

The A + T content of the protein-coding genes, except *NAD3* and *NAD6*, in Zhijin White Goose's mitochondria was higher than 50% (Table 4), and the total content of A + T reached 51.58%, indicating that most protein-coding genes have AT preferences. Since the A + T content of mitochondria of Zhijin White Goose was 52.69%, the non-protein coding regions also have AT preference. The codon usage preference is shown in Table 5. There were 29 preferred codons across 13 PCGs, with 48.28% where the relative synonymous codon usage (RSCU) was greater than 1 for the third base being A/U, and 51.72% where the RSCU was greater than 1 for the third base being C/G. Therefore, no clear preferred choice for the third base of preferred codons was observed. In the mitochondria of Zhijin White Goose, the total probability of the third base of the codon being A/U was 49.42%, and the total probability of it being C/G was 50.58%. These results showed no preference for the third position of mitochondrial codons.

Table 4
The base composition of 13 protein-coding genes in Zhijin White Goose

Gene name	Length	A%	T%	C%	G%	(A + T)%
NAD1	978	27.30	25.46	32.21	15.03	52.76
NAD2	1041	29.88	22.09	35.45	12.58	51.97
CO1	1551	26.95	24.24	32.17	16.63	51.19
CO2	687	28.53	22.13	33.33	16.01	50.66
ATP8	168	34.52	20.24	38.69	6.55	54.76
ATP6	684	29.97	24.12	36.70	9.21	54.09
CO3	784	27.81	22.32	34.06	15.82	50.13
NAD3	351	25.07	24.50	34.19	16.24	49.57
NAD4L	297	25.93	24.92	33.33	15.82	50.84
NAD4	1378	30.12	23.08	34.83	11.97	53.19
NAD5	1818	31.41	21.01	34.38	13.20	52.42
Cytb	1143	28.00	22.57	35.43	14.00	50.57
NAD6	522	9.96	39.66	10.92	39.46	49.62

Table 5
Codon usage frequency of 13 protein-coding genes in Zhijin White Goose

Codon	Count	RSCU	Codon	Count	RSCU	Codon	Count	RSCU	Codon	Count	RSCU
UUU(F)	65	0.56	UCU(S)	26	0.46	UAU(Y)	15	0.28	UGU(C)	4	0.27
UUC(F)	**166**	**1.44**	**UCC(S)**	**89**	**1.57**	**UAC(Y)**	**91**	**1.72**	**UGC(C)**	**26**	**1.73**
UUA(L)	48	0.45	**UCA(S)**	**94**	**1.66**	**UAA(*)**	**6**	**2.18**	**UGA(W)**	**93**	**1.74**
UUG(L)	14	0.13	UCG(S)	18	0.32	UAG(*)	2	0.73	UGG(W)	14	0.26
CUU(L)	44	0.41	CCU(P)	24	0.42	CAU(H)	13	0.24	CGU(R)	2	0.11
CUC(L)	**139**	**1.29**	**CCC(P)**	**84**	**1.46**	**CAC(H)**	**97**	**1.76**	CGC(R)	17	0.93
CUA(L)	**327**	**3.04**	**CCA(P)**	**111**	**1.93**	**CAA(Q)**	**81**	**1.72**	**CGA(R)**	**46**	**2.52**
CUG(L)	73	0.68	CCG(P)	11	0.19	CAG(Q)	13	0.28	CGG(R)	8	0.44
AUU(I)	56	0.40	ACU(T)	38	0.50	AAU(N)	17	0.27	AGU(S)	2	0.07
AUC(I)	**222**	**1.60**	**ACC(T)**	**142**	**1.86**	**AAC(N)**	**108**	**1.73**	**AGC(S)**	**54**	**1.93**
AUA(M)	**137**	**1.53**	**ACA(T)**	**113**	**1.48**	**AAA(K)**	**79**	**1.76**	AGA(*)	1	0.36
AUG(M)	42	0.47	ACG(T)	12	0.16	AAG(K)	11	0.24	AGG(*)	2	0.73
GUU(V)	35	0.73	GCU(A)	40	0.48	GAU(D)	10	0.31	GGU(G)	16	0.29
GUC(V)	**55**	**1.14**	**GCC(A)**	**185**	**2.23**	**GAC(D)**	**55**	**1.69**	**GGC(G)**	**67**	**1.21**
GUA(V)	**77**	**1.60**	**GCA(A)**	**91**	**1.10**	**GAA(E)**	**70**	**1.54**	**GGA(G)**	**97**	**1.75**
GUG(V)	26	0.54	GCG(A)	16	0.19	GAG(E)	21	0.46	GGG(G)	42	0.76

Note: Bold is the preferred codon; RSCU represents the relative synonymous codon usage.

Table 4 here

Table 5 here

The effective number of codons (ENC) method was used to verify the codon preference. When the ENC is ≤ 35, the codon has a significant preference. The ENC of *CO3*, *Cytb*, and *NAD6* in the mitochondria of Zhijin White Goose was ≤ 35 (Fig. 7), indicating that these three PCGs have significant preference. The distribution map of ENC-GC3s is typically used to evaluate the influence of base mutations and natural selection pressure on codon preference. In theory, when there is no selection pressure, ENC and GC3s satisfy the functional relationship ENC = 2 + GC3s + 29/ ($GC3s^2$ +$(1-GC3s)^2$)(Chen *et al.*, 2017). As shown in Fig. 8, the points representing each PCG were distributed below the ENC-GC3s standard curve, indicating that the natural selection pressure impacts the codon preference of Zhijin White Goose PCGs. Among the PCGs, only *ATP8* was close to the ENC-GC3s standard curve, indicating that the codon preference of *ATP8* is more affected by base mutations than other PCGs.

Figure 7 here

Figure 8 here

Results of phylogenetic analysis

The mtGenome sequences of 14 breeds of geese, including Zhijin White Goose, were subjected to multiple sequence alignment and analysis to study the genetic relationships and taxonomic status of Zhijin White Goose and other geese. The

phylogenetic tree is shown in Fig. 9. The phylogenetic tree is composed of two large branches, and the main body of Chinese geese evolution comprises Xupu goose, Zhedong white goose, Yanling goose, Hepu goose, and other geese, with another evolutionary branch of composition. Zhijin white geese and Chinese geese are grouped and are genetically distant from other Chinese geese.

Figure 9 here

Discussion

This study determined the mtGenome of the Zhijin White Goose for the first time. The size of the mtGenome was 16 739 bp. The gene combination was consistent with the typical combination of 37 genes found in the mtGenome of other goose varieties (Lin *et al.*, 2016a, Lin *et al.*, 2016b, Liu, Zhou, Zhang, Luo and Xu, 2013, Ren *et al.*, 2016). The (A + T) content of the mtGenome was 52.69%. The base content of the mtGenome in Zhijin White Goose showed AT preference, which might be affected by three evolutionary forces: mutation, selection, and genetic drift(Kokate *et al.*, 2021).

Due to its unique chemical characteristics and evolutionary status, the RNA secondary structure is closely related to its function. Therefore, RNA secondary structures have been widely used in RNA function and systematic research, and research on molecular evolution and molecular classification has mainly used rRNA. This study used the Auto Traveler software of the RNAcentral website(The, 2019)based on the swan mitochondrial 12S rRNA (b. 16. m. C. melancoryphus) and the human mitochondrial 16S rRNA (mHS_LSU_3D) files to analyse the mitochondria of Zhijin White Goose. The secondary structure of 12S rRNA and 16S rRNA in the genome has been predicted to provide basic data for further research on the structure and function of Zhijin White Goose rRNA.

As an ancient and multifunctional molecule, tRNA contain traces of early life. The secondary structure of the tRNA gene has a splicing signal that can mark certain mtGenome polycistrons (Boore, 1999). The mitochondria of Zhijin White Goose have a standard number of 22 tRNAs, ranging in length from 66 bp to 74 bp. All can form a typical clover structure, except for tRNA-Ser$^{(AGY)}$ due to a missing D arm. Partial mismatches of tRNA genes in the mtGenome can restore gene function through RNA self-shearing without causing amino acid transport obstacles. The mismatches of the Zhijin White Goose are all G-U mismatches, which conform to the G-U swing pairing principle, which is important for maintaining the tRNA secondary structure. Stability is also important. There is a 13-base functional DNA sequence motif (5'-TGGCAGAGCCCGG-3') on the D arm of human mitochondrial tRNA-Leu$^{(UUR)}$, which is the binding site of mitochondrial transcription termination factor (mTERF)(Fernandez-Silva *et al.*, 2003) and is involved in regulating the transcription levels of two rRNA and H-chain downstream genes(Hyvarinen *et al.*, 2007). There is an identical motif (5'- TGGCAGAGCCCGG – 3') on the D arm of the mitochondrial gene tRNA-Leu$^{(UUR)}$ in Zhijin White Goose., indicating that the motif is likely to have similar functions.

The secondary structure of the unique "goose hairpin" sequence(Eberhard *et al.*, 2001) was composed of a stem consisting of seven C/G pairs and a loop containing the TCCC motif. Experiments showed that this motif and the H-chain termination are related(Dufresne *et al.*, 1996). There was a highly similar sequence in the TAS region of Zhijin White Goose. According to the predicted stem-loop structure, the secondary structure was consistent with the characteristics of the "goose hairpin" sequence. This indicated that it might also be related to the termination of the H chain. The key sequences CSB-F, E, D, and C of the CD region of the D-loop region of the Zhijin White Goose are similar to the key sequences of the CSB region of other vertebrates and birds (Cho *et al.*, 2009, Marshall and Baker, 1997, Quinn and Wilson, 1993, Sbisa *et al.*, 1997), and CSB-F is a marker that distinguishes the TAS area from CD, the same as with other geese (Liu, Zhou, Zhang, Luo and Xu, 2013). In the CSB region of Zhijin White Goose, sequences corresponding to mammalian CSB-2 and CSB-3 were not recognised, research has shown that above two sequences among mammals are not universal(Saccone *et al.*, 1991), which might be explained by species specificity (Sbisa, Tanzariello, Reyes, Pesole and Saccone, 1997).

Among the 13 PCGs of the mtGenome, *CO3* and *NAD4* had incomplete stop codons (T-). However, due to the presence of the PolyA tail at the 3'-end of the mRNA, the transcription process automatically completes the TAA stop codon, which does not

affect transcription(Ojala *et al.*, 1981). Codon preference refers to the unequal use of synonymous codons. The main difference between the synonymous codons is the third codon. In this study, the PCGs in Zhijin White Goose mtGenome had an AT preference, display according to RSCU calculation results, the PCGs did not have a clear AT or GC preference. This study used ENC to evaluate codon preference. Only three of out of the thirteen PCGs showed a significant codon preference. By drawing the ENC-GC3 standard curve, it was found that all PCG points were distributed below the standard curve, indicating that the PCG codon preference in Zhijin White Goose is mostly due to the influence of natural selection pressure, and artificial selection is still needed. Studies showed that codon usage preference is related to gene expression strength. Compared with the low-expression genes, the frequency of preference codons used by highly expressed genes had significantly different usage frequencies. Typically, they use a set of preferential synonymous codons. In addition, the preferentially used codons correspond to the most abundant tRNA, thereby improving the translation efficiency(Sabi and Tuller, 2014). In follow-up breeding studies, it is important to strengthen the selection and breeding of high-efficiency expression codons, promote translation efficiency, and increase gene expression to increase the expression of specific proteins and enhance the effect of molecular breeding.

Mitochondrial phylogenetic tree analysis showed that Zhijin White Goose, Swan goose and most Chinese geese belong to an evolutionary branch, in line with the notion that Chinese geese have evolved from swans(Wang *et al.*, 2010). It also proves that the Zhijin White Goose is a local breed with a Chinese goose lineage. However, Sichuan White Goose with the GenBank accession number MK133022 is similar to geese outside China. This might be due to the integrated fragments of nuclear mitochondrial DNA. Mitochondrial DNA is first enclosed in the nucleus and then integrated into the nuclear chromosomes through non-homologous end joining after nuclear double-strand breaks(Hazkani-Covo *et al.*, 2010), resulting in a much slower mutation rate than that of mtGenome. Therefore, this classification appears consistent with previous research results (Ren, Liang, Zhao and He, 2016).

Conclusins

This study measured the Zhijin White Goose mtGenome for the first time. The size was 16 739 bp, the base composition was A (30. 22%), G (15. 08%), C (32.23%), and T (22.47%), and it consisted of 13 PCGs, 22 tRNAs, and 2 rRNAs. All 22 tRNAs, except tRNA-Ser$^{(AGY)}$ have a typical clover structure. Phylogenetic tree analysis based on the mtGenome showed that Zhijin White Goose are closely related to Hongyan and Chinese geese but distantly to geese outside China. The results of this study provide a reference basis for the variety identification of Zhijin White Goose and lay a foundation for the follow-up study of mitochondrial gene function and the development and utilisation of varieties.

Declarations

Acknowledgements

Financial support for this work was obtained (as joint funding) from Guizhou Provincial Science and Technology Plan Project (QKH[2020]1Y047) and (QKH Support [2021] General 156)

Compliance with ethical standards

Conflict of interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Ethical approval

All experimental procedures were conducted under the "Chinese Animal Welfare Ethical Review Laboratory Animal Guidelines" regulations issued by the China Laboratory Animal Standardization Technical Committee (SAC/TC 281) (license

number: GB/T 35 892–2018).

References

1. BENSON G (1999) Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res 27:573–580. doi: 10.1093/nar/27.2.573
2. BOORE JL (1999) Animal mitochondrial genomes. Nucleic Acids Res 27:1767–1780. doi: 10.1093/nar/27.8.1767
3. BOORE JL (2004) Complete mitochondrial genome sequence of Urechis caupo, a representative of the phylum Echiura. BMC Genomics 5:67. doi: 10.1186/1471-2164-5-67
4. CANNONE JJ, SUBRAMANIAN S, SCHNARE MN, D'SOUZA COLLETTJR, DU LM, MULLER YFENGBLINNMADABUSILV, PANDE KM, SHANG N, YU Z, GUTELL RR (2002) The comparative RNA web (CRW) site: an online database of comparative sequence and structure information for ribosomal, intron, and other RNAs. BMC Bioinformatics 3:2. doi: 10.1186/1471-2105-3-2
5. CAO SY, WU XB, YAN P, SU HUYL, JIANG ZG (2006) Complete nucleotide sequences and gene organization of mitochondrial genome of Bufo gargarizans. Mitochondrion 6:186–193. doi: 10.1016/j.mito.2006.07.003
6. CHEN Y, XU Q, TAN C, LI X, CHI X, MA CAIBYUZ, CHEN JL (2017) Genomic analysis of codon usage shows influence of mutation pressure, natural selection, and host features on Senecavirus A evolution. Microb Pathog 112:313–319. doi: 10.1016/j.micpath.2017.09.040
7. CHO HJ, EDA M, NISHIDA S, YASUKOCHI Y, CHONG JR, KOIKE H (2009) Tandem duplication of mitochondrial DNA in the black-faced spoonbill, Platalea minor. Genes Genet Syst 84:297–305. doi: 10.1266/ggs.84.297
8. DARTY K, DENISE A, PONTY Y (2009) VARNA: Interactive drawing and editing of the RNA secondary structure. Bioinformatics 25:1974–1975. doi: 10.1093/bioinformatics/btp250
9. DUFRESNE C, MIGNOTTE F, GUERIDE M (1996) The presence of tandem repeats and the initiation of replication in rabbit mitochondrial DNA. Eur J Biochem 235:593–600. doi: 10.1111/j.1432-1033.1996.00593.x
10. EBERHARD JR, WRIGHT TF, BERMINGHAM E (2001) Duplication and concerted evolution of the mitochondrial control region in the parrot genus Amazona. Mol Biol Evol 18:1330–1342. doi: 10.1093/oxfordjournals.molbev.a003917
11. FERNANDEZ-SILVA P, MONTOYA J (2003) Replication and transcription of mammalian mitochondrial DNA. Exp Physiol 88:41–56. doi: 10.1113/eph8802514
12. HARRISON GL, MCLENACHAN PA, PHILLIPS MJ, SLACK KE, COOPER A, PENNY D (2004) Four new avian mitochondrial genomes help get to basic evolutionary questions in the late cretaceous. Mol Biol Evol 21:974–983. doi: 10.1093/molbev/msh065
13. HAZKANI-COVO E, ZELLER RM, MARTIN W (2010) Molecular poltergeists: mitochondrial DNA copies (numts) in sequenced nuclear genomes. PLoS Genet 6:e1000834. doi: 10.1371/journal.pgen.1000834
14. HYVARINEN AK, POHJOISMAKI JL, REYES A, WANROOIJ S, YASUKAWA T, KARHUNEN PJ, SPELBRINK JN, HOLT IJ, JACOBS HT (2007) The mitochondrial transcription termination factor mTERF modulates replication pausing in human mitochondrial DNA. Nucleic Acids Res 35:6458–6474. doi: 10.1093/nar/gkm676
15. IWASAKI W, FUKUNAGA T, ISAGOZAWA R, YAMADA K, MAEDA Y, SATOH TP, SADO T, MABUCHI K, TAKESHIMA H, MIYA M, NISHIDA M (2013) MitoFish and MitoAnnotator: a mitochondrial genome database of fish with an accurate and automatic annotation pipeline. Mol Biol Evol 30:2531–2540. doi: 10.1093/molbev/mst141
16. JAYAPRAKASH AD, BENSON EK, GONE S, LIANG R, LAMBERTINI SHIMJ, TOLOUE L, WIGLER MM, AARONSON M, SACHIDANANDAM R (2015) Stable heteroplasmy at the single-cell level is facilitated by intercellular exchange of mtDNA. Nucleic Acids Res 43:2177–2187. doi: 10.1093/nar/gkv052
17. KOKATE PP, TECHTMANN SM, WERNER T (2021) Codon usage bias and dinucleotide preference in 29 Drosophila species. *G3 (Bethesda)*, **11**. doi: 10.1093/g3journal/jkab191

18. LASLETT D, CANBACK B (2008) ARWEN: a program to detect tRNA genes in metazoan mitochondrial nucleotide sequences. Bioinformatics 24:172–175. doi: 10.1093/bioinformatics/btm573
19. LAVROV DV, PETT W (2016) Animal Mitochondrial DNA as We Do Not Know It: mt-Genome Organization and Evolution in Nonbilaterian Lineages. Genome Biol Evol 8:2896–2913. doi: 10.1093/gbe/evw195
20. LIN Q, JIANG CAOR, LIU GT, DAI G, ZHANG QZ, HOU SR, HE X (2016a) Determination and analysis of the complete mitochondrial genome sequence of Yanling white goose. Mitochondrial DNA A DNA Mapp Seq Anal 27:1012–1013. doi: 10.3109/19401736.2014.926529
21. LIN Q, JIANG GT, CAO R, YUN L, LI GJ, DAI QZ, ZHANG SR, HOU DX, HE X (2016b) Determination and analysis of the complete mitochondrial genome sequence of Wugangtong grey goose. Mitochondrial DNA A DNA Mapp Seq Anal 27:1008–1009. doi: 10.3109/19401736.2014.926527
22. LIU G, ZHANG ZHOUL, LUO L, XU W (2013) The complete mitochondrial genome of bean goose (Anser fabalis) and implications for anseriformes taxonomy. PLoS ONE 8:e63334. doi: 10.1371/journal.pone.0063334
23. LOWE TM, CHAN PP (2016) tRNAscan-SE On-line: integrating search and context for analysis of transfer RNA genes. Nucleic Acids Res 44:W54–W57. doi: 10.1093/nar/gkw413
24. MARSHALL HD, BAKER AJ (1997) Structural conservation and variation in the mitochondrial control region of fringilline finches (Fringilla spp.) and the greenfinch (Carduelis chloris). Mol Biol Evol 14:173–184. doi: 10.1093/oxfordjournals.molbev.a025750
25. NISHIMAKI T, SATO K (2019) An Extension of the Kimura Two-Parameter Model to the Natural Evolutionary Process. J Mol Evol 87:60–67. doi: 10.1007/s00239-018-9885-1
26. OJALA D, ATTARDI G (1981) tRNA punctuation model of RNA processing in human mitochondria. Nature 290:470–474. doi: 10.1038/290470a0
27. PEREIRA SL, BAKER AJ (2006) A mitogenomic timescale for birds detects variable phylogenetic rates of molecular evolution and refutes the standard molecular clock. Mol Biol Evol 23:1731–1740. doi: 10.1093/molbev/msl038
28. QUINN TW, WILSON AC (1993) Sequence evolution in and around the mitochondrial control region in birds. J Mol Evol 37:417–425. doi: 10.1007/BF00178871
29. RANDI E, LUCCHINI V (1998) Organization and evolution of the mitochondrial DNA control region in the avian genus Alectoris. J Mol Evol 47:449–462. doi: 10.1007/pl00006402
30. REN T, LIANG S, ZHAO A, HE K (2016) Analysis of the complete mitochondrial genome of the Zhedong White goose and characterization of NUMTs: Reveal domestication history of goose in China and Euro. Gene 577:75–81. doi: 10.1016/j.gene.2015.11.018
31. SABI R, TULLER T (2014) Modelling the efficiency of codon-tRNA interactions based on codon usage bias. DNA Res 21:511–526. doi: 10.1093/dnares/dsu017
32. SACCONE C, PESOLE G, SBISA E (1991) The main regulatory region of mammalian mitochondrial DNA: structure-function model and evolutionary pattern. J Mol Evol 33:83–91. doi: 10.1007/BF02100199
33. SBISA E, TANZARIELLO F, PESOLE REYESA, SACCONE C (1997) Mammalian mitochondrial D-loop region structural analysis: identification of new conserved sequences and their functional and evolutionary implications. Gene 205:125–140. doi: 10.1016/s0378-1119(97)00404-6
34. SHARP PM, LI WH (1986) An evolutionary perspective on synonymous codon usage in unicellular organisms. J Mol Evol 24:28–38. doi: 10.1007/BF02099948
35. SLACK KE, DELSUC F, MCLENACHAN PA, ARNASON U, PENNY D (2007) Resolving the root of the avian mitogenomic tree by breaking up long branches. Mol Phylogenet Evol 42:1–13. doi: 10.1016/j.ympev.2006.06.002
36. THE RNAcentral Constortium (2018) A hub of information for non-coding RNA sequences. Nucleic Acids Res 47:D1250–D1251. doi: 10.1093/nar/gky1206

37. WANG CM, WAY TD, CHANG YC, YEN NT, NIEN HUCL, JEA PC, CHEN YS, KAO JY (2010) The origin of the white Roman goose. Biochem Genet 48:938–943. doi: 10.1007/s10528-010-9374-8

38. ZUKER M (2003) Mfold web server for nucleic acid folding and hybridization prediction. Nucleic Acids Res 31:3406–3415. doi: 10.1093/nar/gkg595

Figures

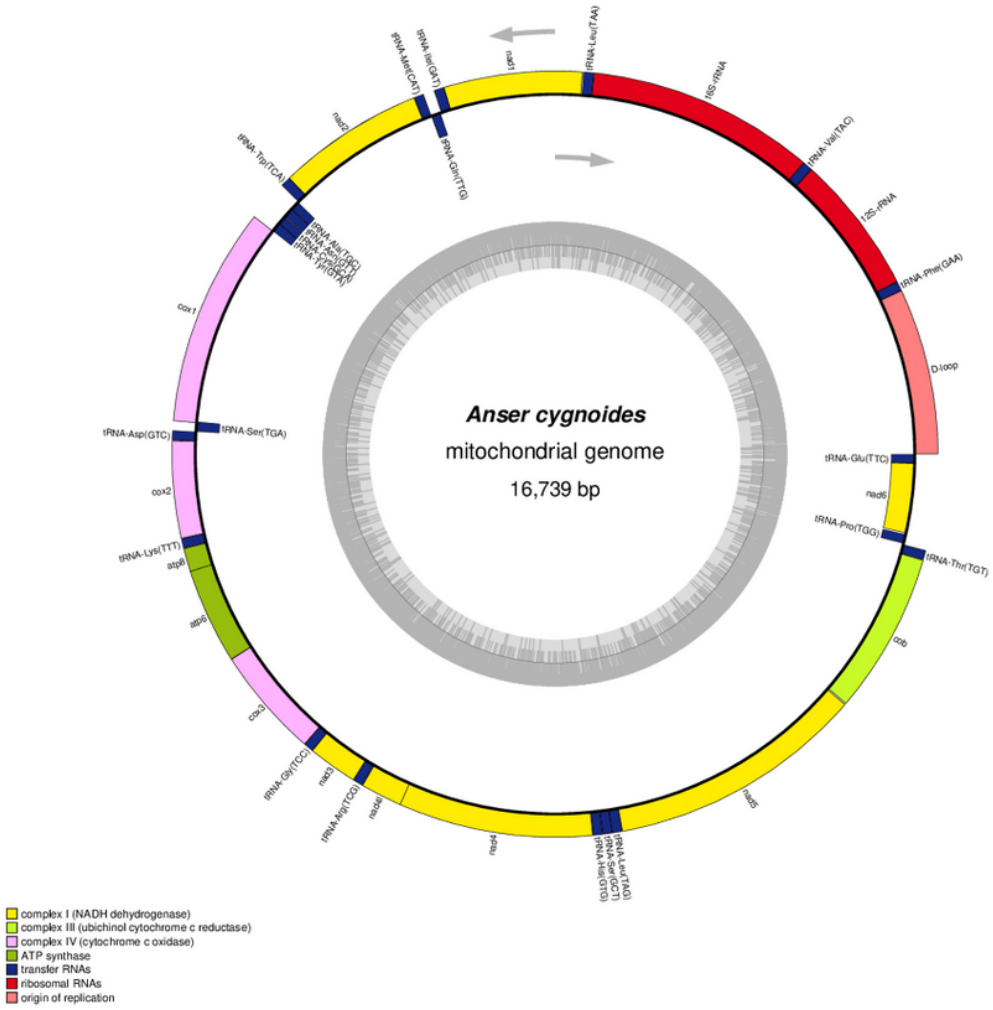

Figure 1

Mitochondrial genome map of Zhijin White Goose

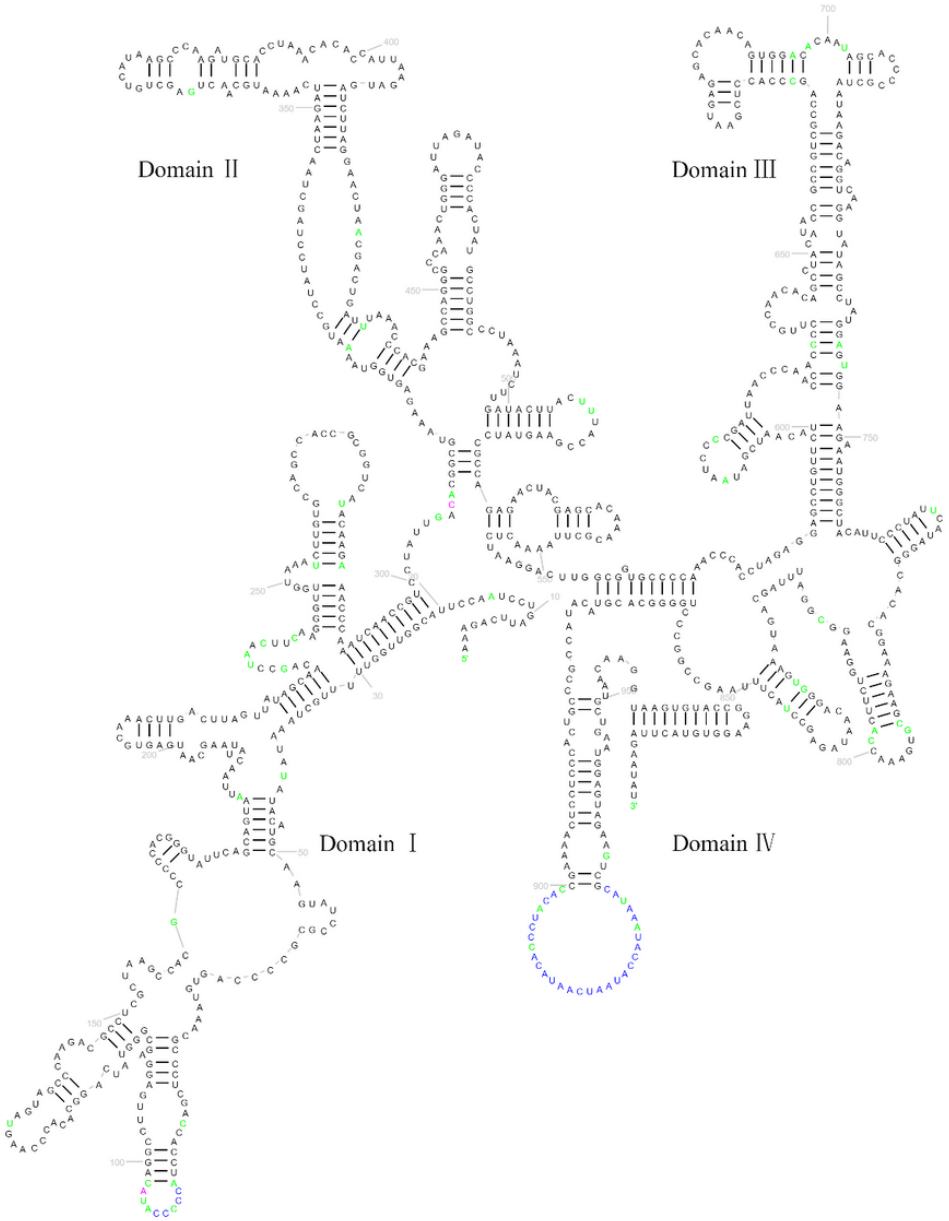

Figure 2

The 12S rRNA secondary structure prediction map of Zhijin White Goose

Note: Black indicates that the base is the same as the template; green indicates that the base is different from the template, red indicates the inserted base, and blue indicates the rearrangement of the base.

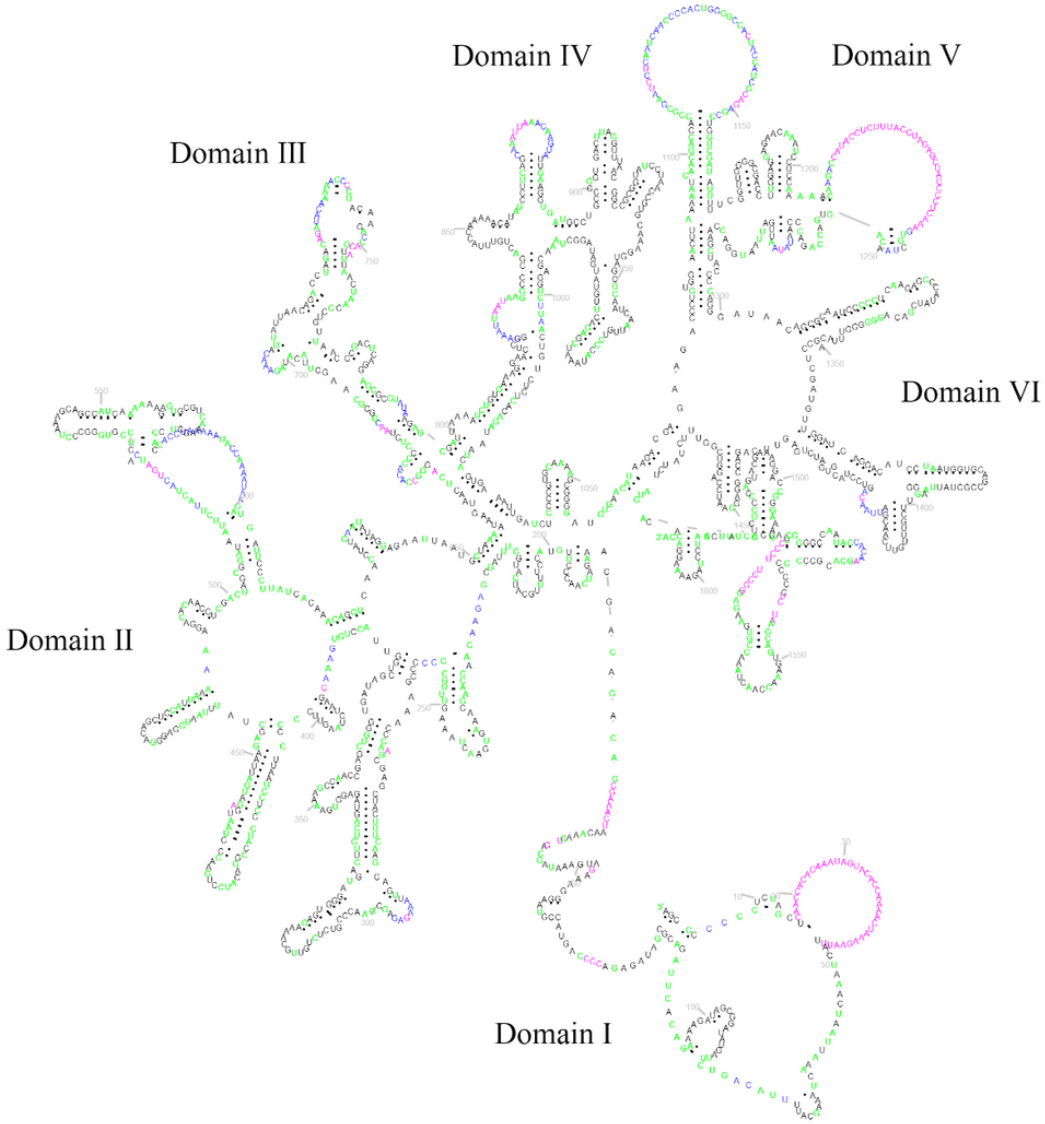

Figure 3

The 16S rRNA secondary structure prediction map of Zhijin White Goose

Note: Black indicates that the base is the same as the template; green indicates that the base is different from the template, red indicates the inserted base, and blue indicates the rearrangement of the base.

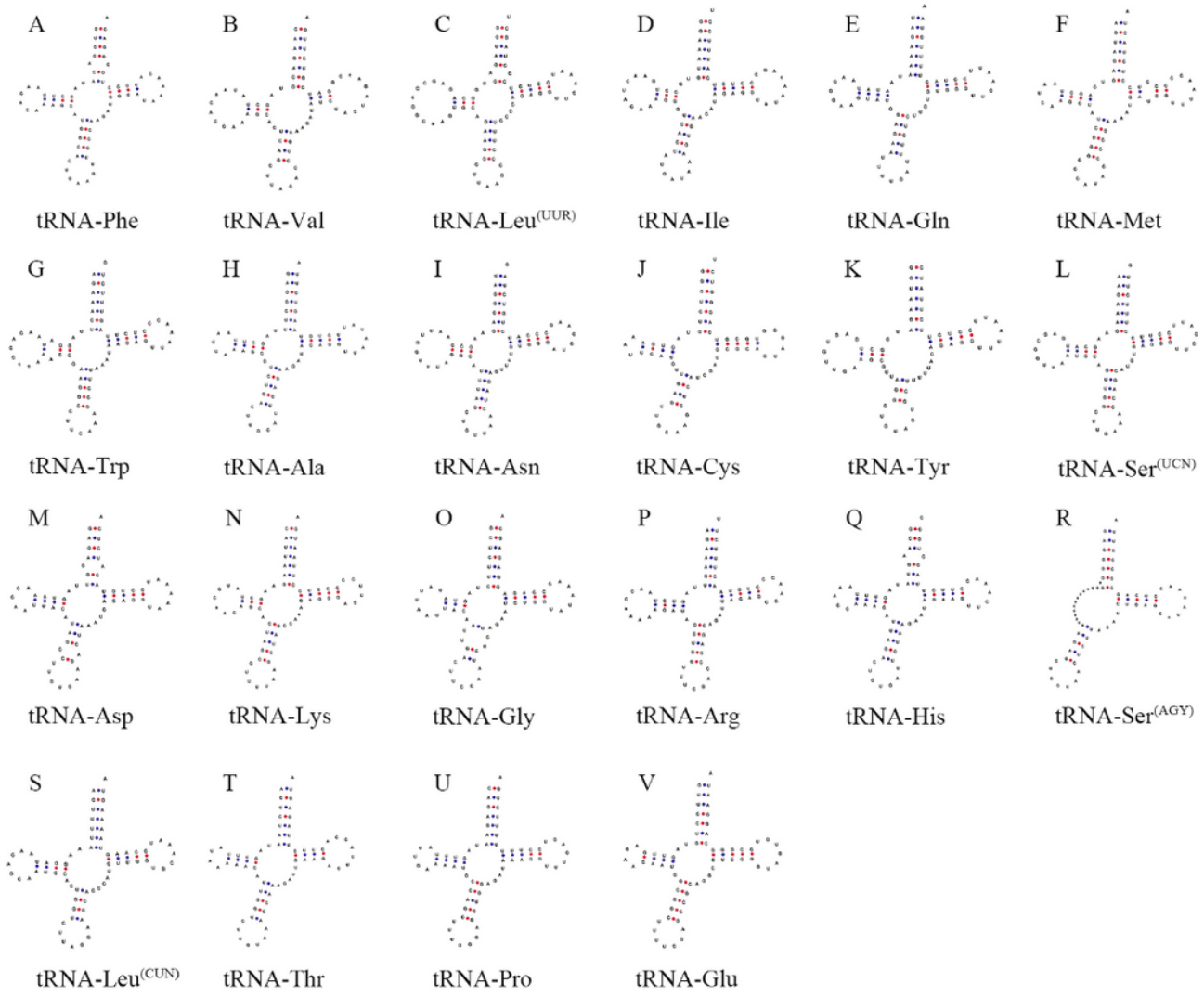

Figure 4

Mitochondrial tRNA secondary structure prediction map of Zhijin White Goose

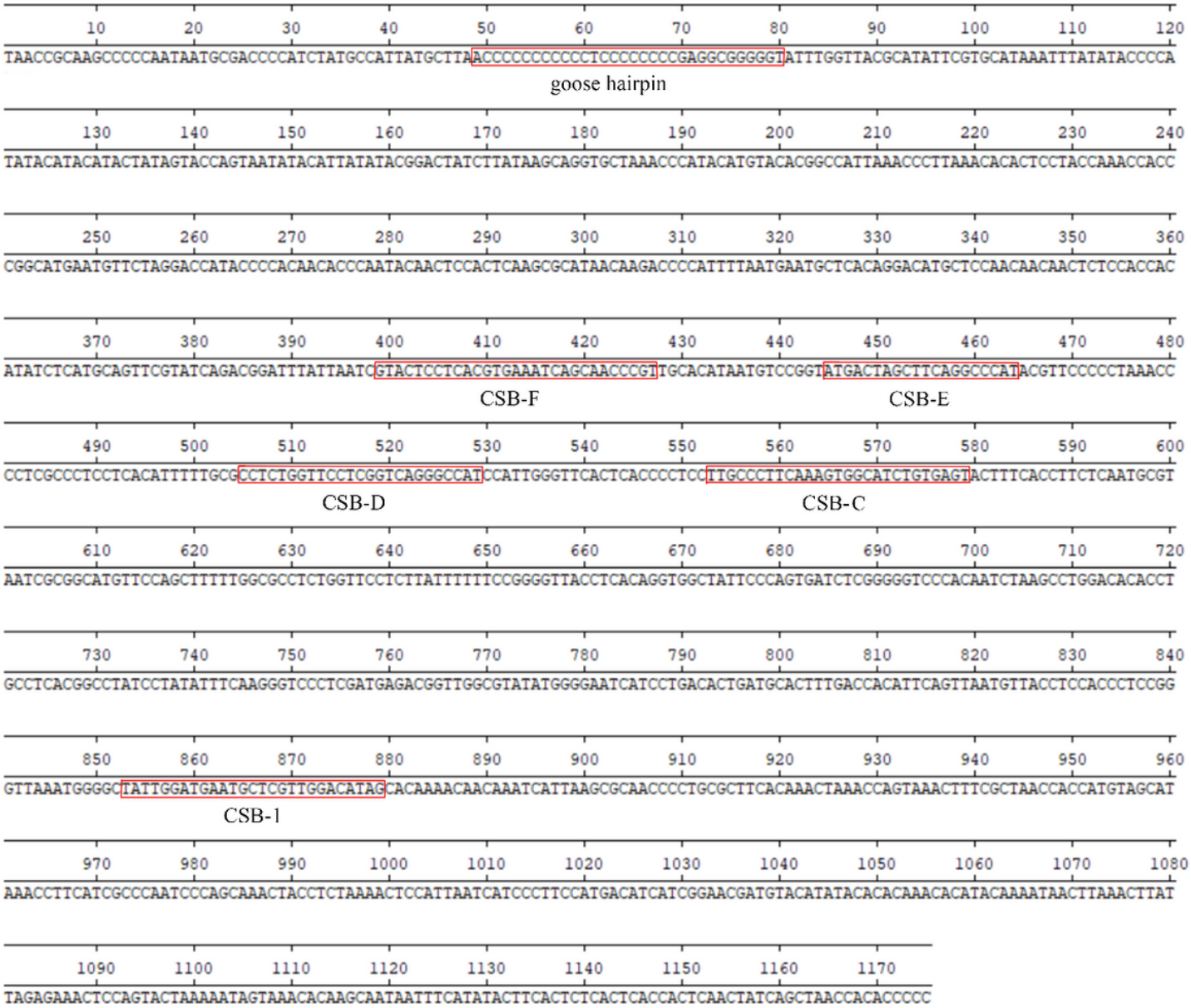

Figure 5

The sequence of the D-loop region of Zhijin White Goose

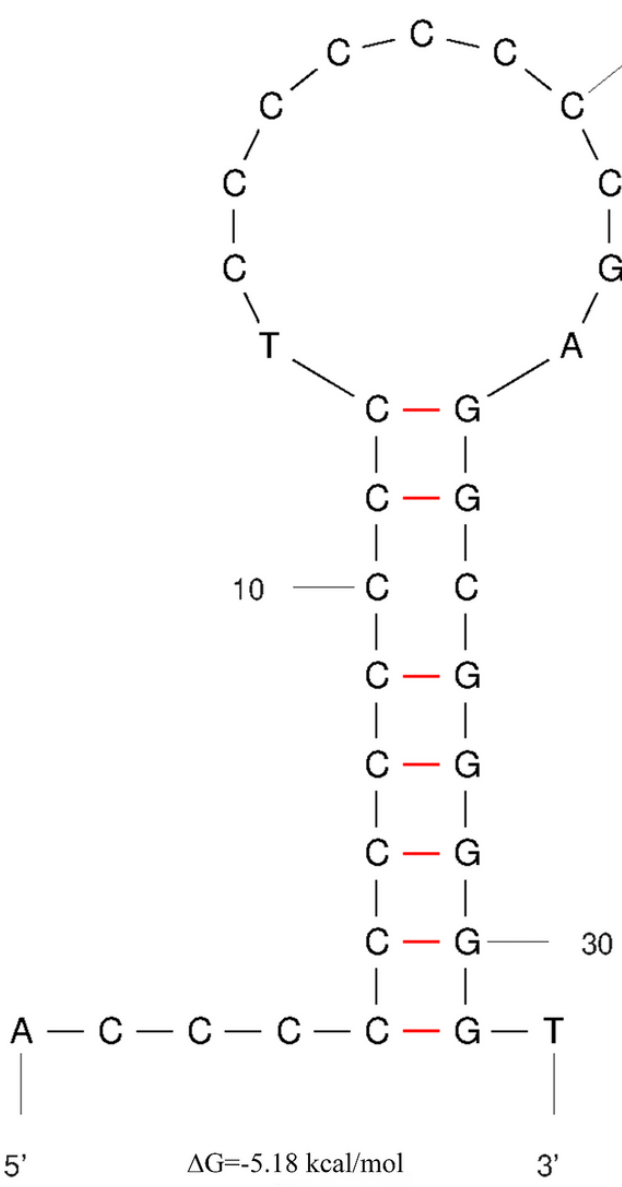

Figure 6

Prediction of the stem-loop structure of the mitochondrial "goose hairpin" sequence of Zhijin White Goose

Note:ΔG represents the change in Gibbs free energy

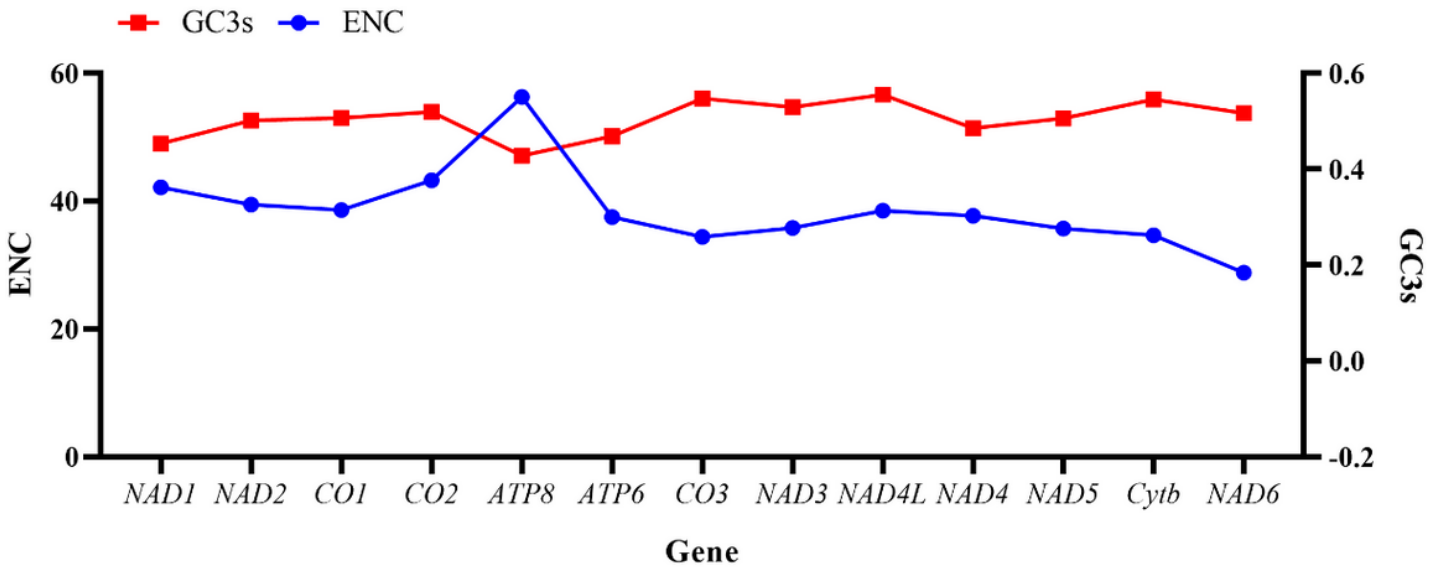

Figure 7

The ENC value and GC3s value of protein-coding genes in the mtGenome of Zhijin White Goose

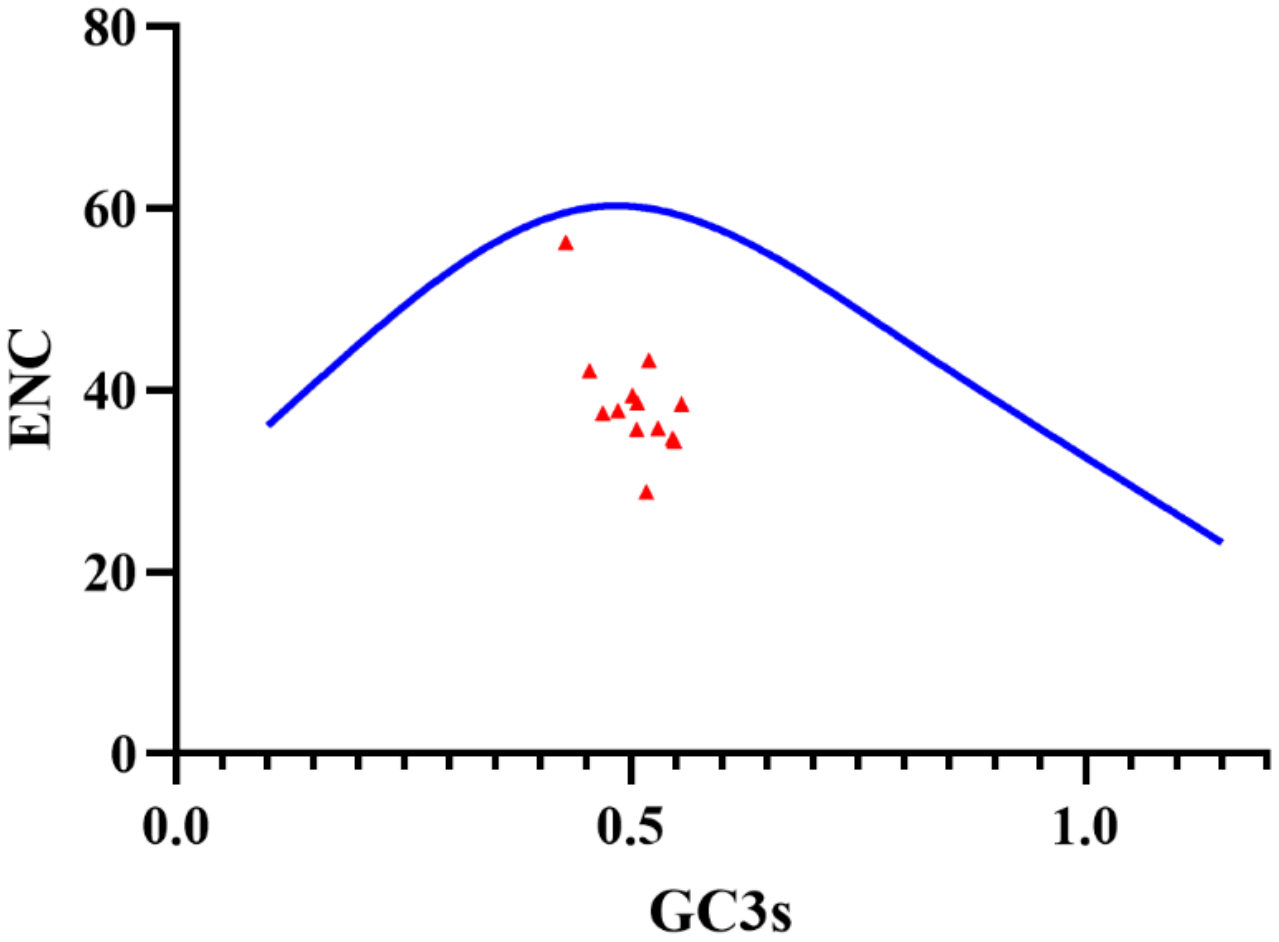

Figure 8

The distribution of ENC-GC3s in the mtGenome of Zhijin White Goose

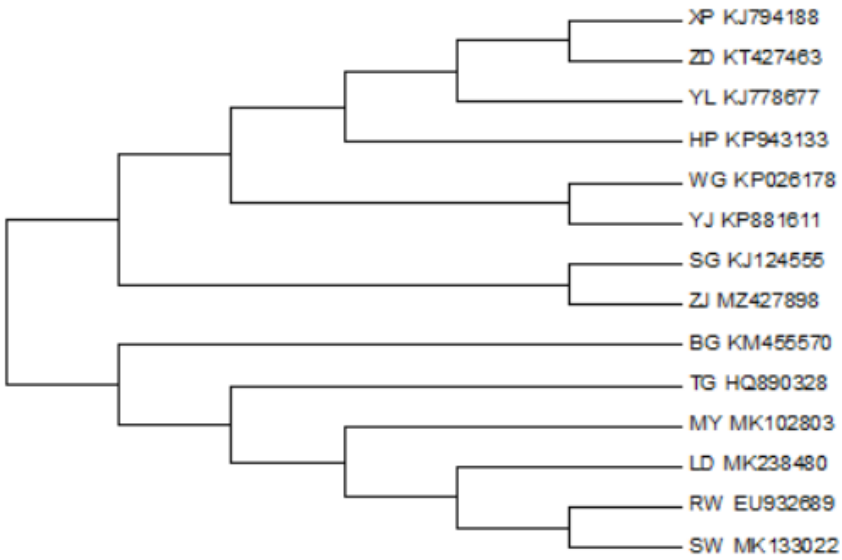

Figure 9

Phylogenetic tree of goose mtGenome based on the neighbor joining method (NJ).

Note:Sichuan White Goose (SW, MK133022), Hepu Goose (HP, KP943133), Youjiang Goose (YJ, KP881611), Wugang Tong Goose(White) (WG, KP026178) , Swan Goose(SG, KJ124555), Landes Goose (LD, MK133021), Roman White Goose (RW, EU932689), Xupu Goose (XP, KJ94188), Zhedong White Goose (ZD, KT427463), Mayang White Goose (MY , MK102803), Bar-headed goose (BG, KM455570), Taiga bean Goose (TG, HQ890328), Yanling Goose (YL, KJ778677)