Structure and variation of the Anseriformes mitochondrial DNA control region

Zuhao Huang and Dianhua Ke

School of Life Sciences, Jinggangshan University, J’ian, Jiangxi Province, P.R. China

Abstract
The control region is the major non-coding segment of animal mitochondrial DNA. To infer the structure and variation of Anseriformes mitochondrial DNA control region, the control region sequences of 52 species were analyzed. The length of the control region sequences ranged from 968 bp (Chenonetta jubata) to 1335 bp (Anseranas semipalmata) and can be separated into three domains. There is a deletion of 100–130 bp in Anatinae, compared to other groups of Anserinae. The average genetic distances among the species within the genera varied from 4.14% (Anser) to 10.58% (Cygnus). The average genetic distances showed insignificantly negative correlation with ts/tv. Domain I is the most variable among the three domains among all the genera. Five conserved sequence boxes in the domain II of Anseriformes sequences were identified. The alignment of the Anseriformes five boxes sequences showed considerable sequence variation. CSB-1, -2 and 3 were not found in the Anseriformes. Maximum-likelihood method was used to construct a phylogenetic tree, which grouped all of the genera into four divergent clades. Anseranas + Chauna and Dendrocygna were identified as early offshoots of the Anatidae. All the remaining taxa fell into two clades that correspond to the two subfamilies Anserinae and Anatinae.

Introduction
The mitochondrial DNA (mtDNA) control region is the major non-coding segment of animal mitochondrial genome. Control region is the most variable part of the vertebrate mtDNA genome, presumably because of the lack of coding constrains (Baker & Marshall, 1997). Because the mutation rates were relatively higher than other mtDNA genes and nuclear genes, the control region sequences were used frequently to estimate phylogenetic relationships (Donne-Gousseé et al., 2002; Huang et al., 2009) and population genetics (Xiao et al., 2013; Zhong et al., 2013) in animals.

In avians, the control region can be divided into three domains: (1) the tRNA^{Glu} -adjacent domain I, (2) the central conserved domain II, (3) the tRNA^{Phe} -adjacent domain III (Donne-Gousseé et al., 2002; Randi & Lucchini, 1998; Ruokonen & Kvist, 2002). Donne-Gousseé et al. (2002) suggested that domain I run from the 5’ end of the control region light strand to position 470, the central domain run from position 471 to 1050, and domain III run from position 1051 to the 3’ end in Anseriformes. The three domains of the control region evolve at high different rates (Randi & Lucchini, 1998). Most of the variability within the control region is concentrated in the domains I and III (Baker & Marshall, 1997). Despite their functional importance, the two peripheral domains have evolved rapidly, generating heterogeneity in both length and base composition (Saccone et al., 1991).

The structure and evolution of avian control region have been analyzed (Baker & Marshall, 1997; Huang & Ke, 2014; Randi & Lucchini, 1998; Ruokonen & Kvist, 2002; Quinn & Wilson, 1993). But only limited number of avian species control region have been studied. Donne-Goussé et al. (2002) preliminary analyzed the structure of Anseriformes control region. But rates and patterns of molecular evolution of Anseriformes control region are not known. The avian order Anseriformes is one of the most important groups of birds. In the present study, we examined the structure of the control region of Anseriformes species based on the complete mitochondrial genome retrieved from GenBank. The aim of this paper is to characterize the structural features and patterns of sequence evolution of the Anseriformes mitochondrial DNA control region.

Materials and methods
All the sequences were collected from the GenBank (species and accession numbers in Table S1). A total of 52 species from 24 genera belonging to the Anseriformes order was analyzed (Table S1).

Sequences were aligned by the MUSCLE (Edgar, 2004). DnaSP v5.0 (Librado & Rozas, 2009) was used to define the variable sites. Nucleotide composition was calculated using MEGA6.0 (Tamura et al., 2013). T-test was used to test the pairwise differences of nucleotide composition between different domains. Sequence divergence among genera was calculated using the Tamura & Nei (1993) model in MEGA 6.0 (Tamura et al., 2013). Only the genera containing at least two species were selected for calculating intrageneric values. Pairwise transition/ transversion (ts/tv) estimates and distribution of variable nucleotide positions were calculated in MEGA6.0 (Tamura et al., 2013).

The relative variability of the domains was tested by calculating the proportion of the variable nucleotides in each domain and
using Chi-square test for proportions to test whether they are significantly different. The boundary of domains I, II and III was defined according to Donne-Goussé et al. (2002). The conserved sequence boxes found were compared with the previously published Alectoris boxes (Randi & Lucchini, 1998) and other avian boxes (Ruokonen & Kvist, 2002).

Modtest 3.0 (Posada & Crandall, 1998) and the Akaiake information criterion (AIC; Posada & Crandall) were used to identify the appropriate nucleotide substitution models. The Maximum likelihood tree (ML; Strimmer & Haeseler, 1996) was obtained using heuristic searches, based on the substitution model proposed by Modtest 3.0 (Posada & Crandall, 1998). We calculated ML tree using MEGA6.0 (Tamura et al., 2013). Statistical support for the internodes in phylogenetic tree was tested by bootstrap percentages (BP) with 1000 replicates (Felsenstein, 1985).

Results and discussion

Alignments

All the Anseriformes species had only one control region. The control region spans the region between the genes for tRNA^{Glu} and tRNA^{Pro} in the Anseriformes species, which was consistent with most of the avian species (Huang & Ke, 2014). But the control region was flanked by the genes for tRNA^{Thr} and tRNA^{Phe} in some species of Picidae, Passeriformes, Falconiformes, and Cuculidae (Bensch & Härld, 2000; Mindell et al., 1998). The length of the control region sequences ranged from 968 bp (Chenonetta jubata) to 1335 bp (Anseranas semipalmata), with an average size of 1109 bp (Table S1). Extensive size variation of the mtDNA control region, attributable to variation in number of tandem repeats, has been reported for many animals (Boyce et al., 1989; Powers et al., 1986; Rand & Harrison, 1989). Within Anatidae, there is a deletion of 100–130 bp in Anatinae (Anas, Mergus, Cairina, Aythya, Aix, Alaprochen, Bucephala, Calonetta, Chloephaga, Clangula, Lophonetta, Marmaronetta, Melanitta, Netta, Somateria), compared to other groups of Anserinae (Cynus, Anser, Dendrocygna, Branta, Cereopsis, Coscoroba), and the divergent genera Chauna (Table S1). Obviously, the deletions are associated with the classification.

Base composition and genetic distance

The average nucleotide composition of Anseriformes control region sequences was: 28.06% A, 25.54% T, 15.41% G, 30.99% C, with a bias against G (Table S2). The amount of A + T was more than that of G + C among whole control region, especially in domain III, same as reported other avian control region (Baker & Marshall, 1997; Ruokonen & Kvist, 2002; Huang & Ke, 2014). The lack of guanines is evident in each domain. Cytosines and adenines are most prevalent in the first domain, thymines and cytosines in the second, and adenines and cytosines in the third (Table S2). Statistically significant differences in nucleotide composition were observed between different domains (p < 0.001), except T between domain I and domain III.

Anseriformes control region sequences were alignable with certainty within genus. Genetic distances between species ranged from 0.14% (between Anser cygnoides and Anser anser) to 46.63% (between Mergus squamatus and Anseranas semipalmata), showing a wide range of divergences. The average genetic distances among the species within the genera varied from 4.14% (Anser) to 10.58% (Cygnus) (Table S3). The average genetic distances showed insignificantly negative correlation with ts/tv (r = -0.368, p > 0.05).

Distribution of variable sites in the control region

Control region are usually considered to be the most variable parts of mtDNA (Randi & Lucchini, 1998). However, position mutability is not distributed randomly across the whole region, but affects particular hypervariable positions and domains (Lytholm et al., 1996; Yang, 1996). The distribution of the variation in the control region seems to be the same in all the species within genera. Average substitution rate for the three domains were 0.85: 0.10: 0.05, corresponding to relative proportions of 85:10:5, respectively. All the genera of Anseriformes, domain I is the most variable of the three domains (Figure 1), with average 85.05% variable positions. Significant difference was observed among seven genera (Chi-square test: Anas, X^2 = 91.72, p < 0.001; Cygnus, X^2 = 8.79, p < 0.05; Anser, X^2 = 91.20, p < 0.001; Mergus, X^2 = 127.76, p < 0.001; Dendrocygna, X^2 = 29.95, p < 0.001; Aythya, X^2 = 103.60, p < 0.001; Branta, X^2 = 33.38, p < 0.001). However, the distribution of the variable positions within domain II of Anseriformes was markedly non-random (Chi-square test, X^2 = 36.09, p < 0.01). The first 200 nucleotides in domain II, apparently evolved at unusually low rates among genera (with only 10.85% variable positions), while the last 200 nucleotides of this domain was hypervariable (with 67.99% variable positions).

Marshall & Baker (1997) believed that mutational hotspots might occur in domain I, whereas more variation was expected in domain III for deeper divergences. Ruokonen & Kvist (2002) also found that variation of some genera was greatest in domain III. Our results did not support this (Figure 1).

---

Figure 1. Distribution of the variable sites in the control region. The number of variable sites within genera has been plotted in 50-bp intervals.
Conserved sequence

Previous comparisons of control region sequences have identified conserved sequence elements based on greater similarity of the sequence elements compared to that of the flanking areas (Ruokonen & Kvist, 2002). We aligned the sequences of the central conserved domain II of every species. Five conserved sequence boxes in the domain II of Anseriformes sequences were localized (Table S4) and identified as boxes F, E, D, C and B. Recently, six central conserved sequence boxes (F to A) were detected in fishes (Gao et al., 2013; Zhang et al., 2011). We did not find the B-box and C-box in *Anseranas semipalmata* and *Chauna torquata*. CSB-1, -2 and 3 were not found in the Anseriformes, as shown for fishes (Xu et al., 2011; Zhang et al., 2011, 2013), avian (Baker & Marshall, 1997), and mammalian species (Walberg & Clayton, 1981).

In C-box, 20 of 25 nucleotide positions were fully conserved among the Anseriformes sequences. While, there were three nucleotide positions were variable in E-box and D-box, eight in F-box and B-box (Table S4).

Phylogenies

On the basis of hierarchical likelihood-ratio tests (hLRTs) as implemented in Modeltest 3.0 (Posada & Crandall, 1998), the model Hasegawa–Kishino–Yano (HKY) model (Hasegawa et al., 1985) + Gamma distribution was used (HKY + G, -lnL = 9121.227, AIC = 4103.101). The ML phylogenetic tree was estimated with the best-fit model HKY + G based on the mtDNA control-region of Anseriformes. The phylogenetic tree grouped all of the genera into four divergent clades (A, B, C and D, Figure 2). *Chauna + Anseranas* was the first to split from the Anseriformes lineage (BP = 100%). Monophyly of the *Dendrocygna* clade was well supported with high bootstrap values (BP = 100%). Analysis of control-region genes supported the others genera fell into two clades corresponding to the two subfamilies Anserinae (C, BP = 97%) and Anatiane (D, BP = 100%).

The phylogenetic tree supported a basal position of *Chauna + Anseranas* in Anseriformes (Figure 2). Traditionally, members of *Chauna* were considered to be Anhimidae and *Anseranas* belonged to Anatidae based on morphological and behavioral characters (Delacour & Mayr, 1945; Del Hoyo et al., 1992; Johnsgard, 1978). The molecular phylogeny challenges the traditional classification of the Anseriformes (Donne-Gousse et al., 2002; Sibley & Ahlquist, 1990;). Sibley & Ahlquist (1990) first attempted to clarify the phylogeny of the Anseriformes using molecular data. Sibley & Ahlquist (1990) considered that *Chauna + Anseranas* clustered together and formed the suborder Anhimidae based on DNA hybridization. Control-region genes data well supported the result of Sibley & Ahlquist (1990). Our molecular results suggested that the genus *Dendrocygna* diverged from other Anatidae earlier than the Anatinae/Anserinae split. Within Anatidae except *Dendrocygna*, control-region genes supported the conventional division between Anatianes and Anserinae. This basal dichotomy was also observed by other molecular studies (Donne-Gousse et al., 2002; Huang et al., 2014; Sorenson et al., 1999).

Declaration of interest

The authors report no conflicts of interest. The authors alone are responsible for the content and writing of the paper. This work was supported by the National Natural Science Foundation of China (30960051, 31260088), Jiangxi Province Talent Project 555, Jiangxi Province Major Disciplines Academic Leaders (20133BCB22010), and the Natural Science Foundation of Jiangxi Province (20132BAB204022).

References

Baker AJ, Marshall HD. (1997). Mitochondrial control region sequences as tools for understanding evolution. In: Mindell DP, editor. Avian molecular evolution and systematics. San Diego: Academic Press. p. 51–82.

Bensch S, Härlid A. (2000). Mitochondrial genomic rearrangements in songbirds. Mol Biol Evol 17:107–13.

Boyce TM, Zwick ME, Aquadro CE. (1989). Mitochondrial DNA in the bark weevils: Size, structure and heteroplasmy. Genetics 123:825–36.

Delacour J, Mayr E. (1945). The family Anatidae. Wilson Bull 57:3–55.

Del Hoyo J, Elliott A, Sargatal J. (1992). Handbook of the birds of the world. Vol I. Barcelona: Lynx Editions.
Donne-Gousse C, Laudet V, Hänni C. (2002). A molecular phylogeny of anseriformes based on mitochondrial DNA analysis. Mol Phylogenet Evol 23:339–56.

Edgar RC. (2004). MUSCLE: Multiple sequence alignment with high accuracy and high throughput. Nucleic Acid Res 32:1792–97.

Felsenstein J. (1985). Confidence limits on phylogenies: An approach using the bootstrap. Evolution 39:783–91.

Gao TX, Li N, Zhang YP, Shi P. (2013). The complete mitochondrial genome of Japanese sandeel Ammodytes personatus (Perciformes, Ammodytidae): Rare structure in mitochondrial DNA compared. Mitochondrial DNA 24:320–2.

Hasegawa M, Kishino H, Yano T. (1985). Dating of the human-ape split by a molecular clock of mitochondrial DNA. J Mol Evol 22:160–74.

Huang ZH, Ke DH. (2014). Structure and evolution of the Phasianidae mitochondrial DNA control region. Mitochondrial DNA doi:10.3109/19401736.2014.895987.

Huang ZH, Yang CZ, Ke DH. (2014). DNA barcoding and phylogenetic relationships in Anatidae. Mitochondrial DNA DOI: 10.3109/19401736.2014.926545.

Huang ZH, Liu NF, Xiao YA, Chen YL, Mei WF, Wen LY, Zhang LX, Yu XP. (2009). Phylogenetic relationships of four endemic genera of the Phasianidae in China based on mitochondrial DNA control region genes. Mol Phylogenet Evol 52:378–83.

Johnsgard PA. (1978). Ducks, geese, and swans of the world. Lincoln: University of Nebraska Press.

Librado P, Rozas J. (2009). Dnasp v5: A software for comprehensive analysis of DNA polymorphism data. Bioinformatics 25:1451–2.

Lytholm T, Leimar O, Gyllensten U. (1996). Low diversity and biased substitution patterns in the mitochondrial DNA control region of sperm whales: Implications for the estimates of time since common ancestry. Mol Biol Evol 13:1318–26.

Mindell DP, Sorenson MD, Dimcheff DE. (1998). Multiple independent origins of mitochondrial gene order in birds. Proc Natl Acad Sci USA 95:10693–7.

Posada D, Buckley TR. (2004). Model selection and model averaging in phylogenetics: Advantages of akaike information criterion and Bayesian approaches over likelihood ratio tests. Syst Biol 53:793–808.

Posada D, Crandall KA. (1998). Modeltest: Testing the model of DNA substitution. Bioinformatics 14: 817–8.

Powers TO, Platzer EG, Hyman C. (1986). Large mitochondrial genome and mitochondrial DNA size polymorphism in the mosquito parasite, Romanomermis culicivorax. Curr Genet 11:71–7.

Quinn TW, Wilson AC. (1993). Sequence evolution in and around the mitochondrial control region in birds. J Mol Evol 37:417–25.

Rand DM, Harrison RG. (1989). Molecular population genetics of mtDNA size variation in crickets. Genetics 121:551–69.

Randi E, Lucchini V. (1998). Organization and evolution of the mitochondrial DNA control region in the avian genus Alectoris. J Mol Evol 47:449–62.

Ruokonen M, Kivist L. (2002). Structure and evolution of the avian mitochondrial control region. Mol Phylogenet Evol 23:422–32.

Saccone C, De Giorgi C, Gissi C, Pesole G, Reyes A. (1999). Evolutionary genomics in Metazoa: The mitochondrial DNA as a model system. Gene 238:195–209.

Sibley CG, Ahlquist JE. (1990). Phylogeny and classification of birds: A study in molecular evolution. New Haven, CT: Yale University Press.

Sorenson MD, Cooper A, Paxinos EE, Quinn TW, James HF, Olson SL, Fleischer R. (1999). Relationships of the extinct moa-nalos, flightless Hawaiian waterfowl, based on ancient DNA. Proc R Soc Lond 266: 2187–93.

Stamatakis A, Hoover P, Rougemont J. (2008). A rapid bootstrap algorithm for the RAxMLweb servers. Syst Biol 57:91–8.

Tamura K, Nei M. (1993). Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. Mol Biol Evol 10:512–26.

Tamura K, Stecher G, Peterson D, Filipski A, and Kumar S. (2013). MEGA6: Molecular evolutionary genetics analysis version 6.0. Mol Biol Evol 30:2725–9.

Walberg MW, Clayton DA. (1981). Sequence and properties of the human KB cell and mouse L cell D-loop regions of mitochondrial DNA. Nucleic Acids Res 9:5411–21.

Xiao YS, Song N, Li J, Xiao ZZ, Gao TX. (2013). Significant population genetic structure detected in the small yellow croaker Larimichthys polyactis inferred from mitochondrial control region. Mitochondrial DNA doi:10.3109/19401736.2013.843076.

Xu TJ, Cheng YZ, Sun YN, Shi G, Wang RX. (2011). The complete mitochondrial genome of bighead croaker, Collichthys nasus (Perciformes, Sciaenidae): Structure of control region and phylogenetic considerations. Mol Biol Rep 38:4673–85.

Yang Z. (1996). Among-site rate variation and its impact on phylogenetic analyses. Trends Ecol Evol 11:367–72.

Zhang Y, Zhang H, Gao TX, Miao ZQ. (2011). Structure of mitochondrial DNA control region and molecular phylogenetic relationship among three flounders of genus Pleuronectes. Bioch Syst Ecol 39:627–34.

Zhong LQ, Song C, Wang MH, Chen YM, Qin Q, Pan JL, Chen XH. (2013). Genetic diversity and population structure of yellow catfish Pelteobagrus fulvidraco from five lakes in the middle and lower reaches of the Yangtze River, China, based on mitochondrial DNA control region. Mitochondrial DNA 24:552–8.