MOLECULAR ECOLOGY

Molecular Ecology (2014) **23**, 5393–5397

NEWS AND VIEWS

OPINION

Mitochondrial genomes of domestic animals need scrutiny

NI-NI SHI,*†[1] LONG FAN,‡[1] YONG-GANG YAO,† § MIN-SHENG PENG*† and YA-PING ZHANG*†¶

**State Key Laboratory of Genetic Resources and Evolution, and Yunnan Laboratory of Molecular Biology of Domestic Animals, Kunming Institute of Zoology, Chinese Academy of Sciences, Kunming, Yunnan 650223, China; †Kunming College of Life Science, University of Chinese Academy of Sciences, Kunming, Yunnan 650204, China; ‡School of Life Sciences, The Chinese University of Hong Kong, Shatin, Hong Kong, SAR 999077, China; §Key Laboratory of Animal Models and Human Disease Mechanisms, Kunming Institute of Zoology, Chinese Academy of Sciences, Kunming, Yunnan 650223, China; ¶Laboratory for Conservation and Utilization of Bio-Resources & Key Laboratory for Microbial Resources of the Ministry of Education, Yunnan University, Kunming, Yunnan 650091, China*

More than 1000 complete or near-complete mitochondrial DNA (mtDNA) sequences have been deposited in GenBank for eight common domestic animals (cattle, dog, goat, horse, pig, sheep, yak and chicken) and their close wild ancestors or relatives, as well. Nevertheless, few efforts have been performed to evaluate the sequence data quality. Herein, we conducted a phylogenetic survey of these complete or near-complete mtDNA sequences based on mtDNA haplogroup trees for the eight animals. We show that errors due to artificial recombination, surplus of mutations and phantom mutations do exist in 14.5% (194/1342) of mtDNA sequences and all of them should be treated with wide caution. We propose some caveats for future mtDNA studies of domestic animals.

Keywords: data quality, domestic animal, haplogroup, mtDNA, phylogeny

Received 16 July 2014; revised 22 September 2014; accepted 03 October 2014

Mitochondrial DNA (mtDNA) is one of the most widely used markers in exploring genetic diversity and tracing evolutionary history for human, domestic and wild animals. Since 2006, we witnessed the change from partial mtDNA sequence (e.g. control region) to complete mitochondrial genome in abundant studies of various domestic animals (Wang *et al.* 2014). Hitherto more than 1000 complete or near-complete mtDNA sequences have been deposited in GenBank for eight common domestic animals (cattle, dog, goat, horse, pig, sheep, yak and chicken) and their close wild ancestors or relatives (Table 1; Table S1, Supporting information). However, most researchers took those reported mtDNA genomes in their data mining and comparative analyses without any scrutiny for data quality. Unfortunately, it is known that some sequences have been shown to contain sequencing errors (Achilli *et al.* 2012; Miao *et al.* 2013) and nuclear mtDNA (NUMT) contaminations (Hassanin *et al.* 2010), similar to problems found in human mtDNA studies (Yao *et al.* 2008, 2009).

Most released mtDNA genome sequences of domestic animals were generated through several PCRs and Sanger sequencing reactions (e.g. Pang *et al.* 2009; Yu *et al.* 2013). These practices are labour-intensive and prone to errors, as has been well demonstrated in human mtDNA data generated through similar experimental protocols (Bandelt *et al.* 2006). It triggers us to ask: do similar errors also occur in the mitochondrial genome data of domestic animals? Herein, we summarized three major kinds of errors (Table 2) and screened these errors in 1342 complete or near-complete mtDNA sequences from eight domestic animals and their close wild ancestors or relatives (Table 1; Table S1, Supporting information).

We adopted the phylogenetic strategy developed in human mtDNA analyses (Yao *et al.* 2009) based on high-resolution genealogy (i.e. mtDNA haplogroup tree). In brief, mtDNA haplogroup tree can be constructed with the parsimony-like methods such as networks (Bandelt *et al.* 1999). The haplogroups can be defined by following a specific diagnostic mutational motif. Specifically, the members that belong to a certain haplogroup can be characterized by a string of mutations that define the internal branch. This branch directs to the haplogroup (internal node) in the tree (van Oven & Kayser 2009). When comparing with the defined reference sequence (Bandelt *et al.* 2014), variants in each sequence can be scored. According to the haplogroup tree, the variants can be mapped on the internal branch as diagnostic mutations or on the terminal branch (tip). As a result, sequences with anomalous variants showing conflicts with their known phylogenetic status can be identified and shall be treated with caution. This phylogenetic method has been proved to be powerful and sensitive for human mtDNA data quality assessment (Bandelt *et al.* 2005, 2006; Salas *et al.* 2005a; Kong *et al.* 2008; Yao *et al.* 2009).

Since 2007, the mtDNA haplogroup trees for pig (Wu *et al.* 2007), cattle (Achilli *et al.* 2008, 2009; Bonfiglio *et al.* 2010, 2012), horse (Achilli *et al.* 2012), chicken (Miao *et al.* 2013) and sheep (Lancioni *et al.* 2013) have been constructed. In

Correspondence: Min-Sheng Peng, Fax: +86-871-6519-5430; E-mail: pengminsheng@mail.kiz.ac.cn or Ya-Ping Zhang, Fax: +86 871-6519 5430; E-mail: zhangyp@mail.kiz.ac.cn

[1] These authors contributed equally to this work.

Table 1 Summary of (near-)complete mtDNA genome sequences analysed in this work

Animals	Related species	Released by GenBank*	Reference sequence	Haplogroup nomenclature
Cattle and Aurochs	*Bos taurus*	266	V00654	I, P, Q, R and T
	B. indicus	10		
	B. primigenius	2		
Yak and wild yak	*B. grunniens*	79	GQ464259[†]	A – D
Dog and grey wolf	*Canis lupus*	447	EU789787[†]	A – F
Horse and Przewalski's horse	*Equus caballus*	254	JN398377	A – R
	E. przewalskii	9		
Chicken and Red Junglefowl	*Gallus gallus*	66	AP003321	A – I and X – Z
Pig and wild boar	*Sus scrofa*	127	EF545567	A, D and E
Sheep	*Ovis aries*	47	AF010406	A – E
Goat	*Capra hircus*	35	GU068049	A – C

*Access time: 1 June 2014.

[†]Reference sequences are defined in this work.

Table 2 Three kinds of common errors occurring in mtDNA data analysed in this study

Errors	Common phenotypes	Major causes
Artificial recombination	Missing diagnostic mutations; mis-added diagnostic mutations of different haplogroups	Sample mix-up; contamination
Surplus of mutations	Excessive unusual mutations, especially transversions, indels and heteroplasmic mutations	Sequencing errors; NUMTs contamination
Phantom mutations	Mutations are laboratory specific and occur repeatedly on different haplogroup backgrounds	Low quality of sequencing; technical pitfalls of NGS

Table 3 mtDNA genome sequences with potential errors in eight domestic animals

Animals	Released by GenBank	Artificial recombination	Surplus of mutations	Phantom mutations	Percentage of potential errors, %
Cattle	278	1	1	7 + 34*	15.46
Chicken	66		1		1.51
Dog	447	11	5		3.57
Goat	35		4		11.42
Horse	263	20	2	1 + 71*	35.74
Pig	127	7	19	15	32.28
Sheep	47		2		4.25
Yak	79	7	1		10.12

*Detected in next-generation sequencing data (Table S6, Supporting information).

terms of the available trees, we followed the proposed caveats (Bandelt *et al.* 2005, 2006; Kong *et al.* 2008; Yao *et al.* 2008, 2009) to investigate the data quality. The previously defined reference sequences (Table 1) were adopted as the consistent standards to avoid confusion and misunderstanding (Salas *et al.* 2012; Bandelt *et al.* 2014). We discovered that potential errors occurred in cattle, horse, pig, sheep and chicken (Table 3; Table S2, Supporting information). For instance, in five pig mitochondrial genomes published recently (Yu *et al.* 2013), two sequences (GenBank Accession nos. KC250273 and KC469586) were problematic (Fig. 1a). The mis-added variants C9553T-C9605T in KC250273 (which belongs to haplogroup D1e) are the diagnostic motif of haplogroup E1a. Similarly, motifs T2374C (i.e. @2374)-T2613C, A5672G-C5678T-T5708C-T5753C, C9156T-C9157T-C9225T-T9234C and T14812C-C14839T in KC469586 (which belongs to haplogroup D1a1a) characterize haplogroup E. The missed and mis-added mutations most likely represent as 'recombinants' of separate segments from different samples of haplogroups E1a and E that are common in European pigs. The errors of artificial recombination are probably due to sample mix-up or contamination with European pig breed(s).

We followed the same approach to check mitochondrial genome data from other domestic animals. We first reconstructed mtDNA haplogroup trees for dog, goat and

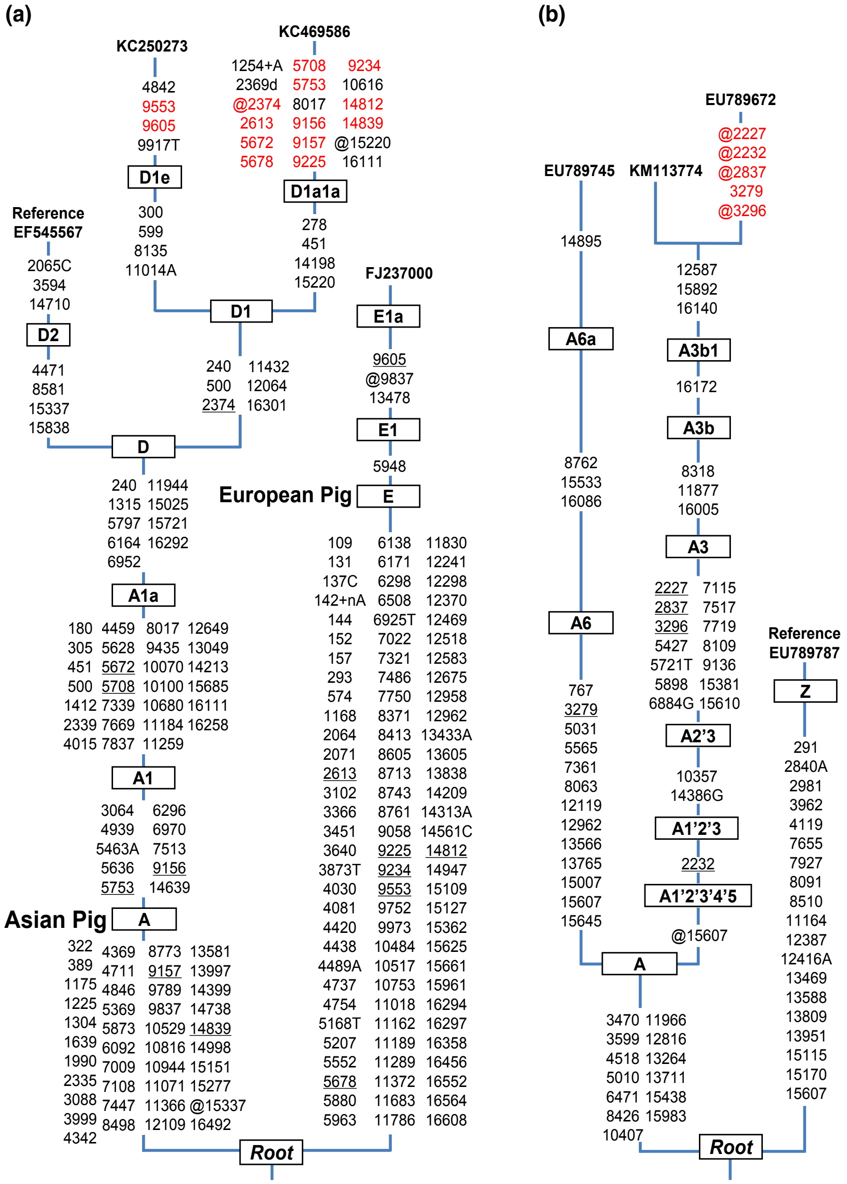

Fig. 1 The flawed mitochondrial genome sequences in pigs (a) and dogs (b). The nucleotide positions in the sequences from pigs and dogs were scored relative to the reference sequences EF545567 and EU789787, respectively. Transitions are shown on the branches, and transversions are further annotated by adding suffixes. Deletions and insertions are indicated by 'd' and '+', respectively. Mutations towards a base identical-by-state to the reference sequences are indicated with the prefix @. Suspected mutations due to errors are in red, and their corresponding phylogenetic status in haplogroup trees are underlined.

yak (Tables S3–S5, Supporting information) based on the parsimony-like method; bibliographically, this method is usually applied in domestic animals (e.g. Wu *et al.* 2007; Achilli *et al.* 2008; Miao *et al.* 2013) as well as in human (van Oven & Kayser 2009) mtDNA studies. The hierarchical haplogroup nomenclature systems were updated from previous studies (Pang *et al.* 2009; Wang *et al.* 2010; Doro *et al.* 2014), referring to the rules in human mtDNA phylogeny (van Oven & Kayser 2009). Not unexpectedly, several potential errors are detected in the mtDNA sequences of these animals (Table 3; Table S2, Supporting information). For instance, a dog sequence (EU789672) from our previous study (Pang *et al.* 2009) was suspected to be corroded by artificial recombination. The diagnostic variants G2232A and T2227C-T2837C-G3296A, which define haplogroups A1′2′3 and A3, respectively, are missing (Fig. 1b). It was probably due to sample mix-up after replacing the fragment around nucleotide position 2200–3300 of EU789672 with another sample of nonhaplogroup A1′2′3 (most likely haplogroup A6, Fig. 1b). We re-amplified this mtDNA and sequenced the entire mitochondrial genome. Indeed, the suspected artificial recombination is confirmed (Fig. 1b). The corrected sequence has been updated in GenBank as KM113774.

In addition to artificial recombination, we show that phantom mutation and poor sequencing quality do exist in mtDNA data of domestic animals (Table 3; Table S2, Supporting information). When using the flawed sequences, it is prone to getting erroneous claims. In some cases, it is expected that the surplus of mutations can 'increase' length of branch in the phylogeny and 'accelerate' molecular evolution (e.g. detection of positive selection and violation of molecular clock). Indeed, there are several cases reported in human mtDNA studies (e.g. Salas *et al.* 2005b; Liu *et al.*

2012). Moreover, phantom mutations basically are detected due to technological pitfalls in next-generation sequencing platforms (especially for homopolymers) (Table S6, Supporting information). Similarly, data qualities of some human mtDNA sequences generated via next-generation sequencing platforms were less than satisfactory (Bandelt & Salas 2012).

Along with the progress of sequencing techniques, accumulation of mitochondrial genome sequences from domestic animals is accelerating. The common practice of generating and analysing mtDNA data should be carried out with the necessary caution. Analyses based on traditional phylogenetic software, at least in the related literatures (e.g. Pang *et al.* 2009; Yu *et al.* 2013), are inefficient to discern those errors. Thus, we suggest researchers to follow some caveats from human mtDNA studies during experiments and data analyses (Bandelt *et al.* 2005, 2006; Salas *et al.* 2005a; Kong *et al.* 2008; Yao *et al.* 2008, 2009). Phylogenetic analyses based on mtDNA haplogroup trees are recommended, although it may be difficult for a nonexpert or a novice in mtDNA analysis. Developing bioinformatic tools to make the related analyses convenient and efficient should be encouraged in the future. Furthermore, we suggest that these flawed sequences identified in this study (Table S2, Supporting information) should be never used in future data-mining analyses, unless a correct one was updated. We welcome the researchers of the original study to take our warnings into consideration and double check these sequences tagged with ‘flawed’, just as what we did for the dog sequence EU789672.

Acknowledgements

We thank Adeniyi C. Adeola for technical assistance. This work was supported by the 973 programme (2013CB835200, 2013CB835204). The Youth Innovation Promotion Association, Chinese Academy of Sciences (CAS), provided support to M-S.P. M-S.P. is a member of Youth Innovation Promotion Association, CAS.

References

Achilli A, Olivieri A, Pellecchia M *et al.* (2008) Mitochondrial genomes of extinct aurochs survive in domestic cattle. *Current Biology*, **18**, R157–R158.

Achilli A, Bonfiglio S, Olivieri A *et al.* (2009) The multifaceted origin of taurine cattle reflected by the mitochondrial genome. *PLoS ONE*, **4**, e5753.

Achilli A, Olivieri A, Soares P *et al.* (2012) Mitochondrial genomes from modern horses reveal the major haplogroups that underwent domestication. *Proceedings of the National Academy of Sciences of the United States of America*, **109**, 2449–2454.

Bandelt HJ, Salas A (2012) Current next generation sequencing technology may not meet forensic standards. *Forensic Science International Genetics*, **6**, 143–145.

Bandelt HJ, Forster P, Röhl A (1999) Median-joining networks for inferring intraspecific phylogenies. *Molecular Biology and Evolution*, **16**, 37–48.

Bandelt HJ, Kong QP, Parson W, Salas A (2005) More evidence for non-maternal inheritance of mitochondrial DNA? *Journal of Medical Genetics*, **42**, 957–960.

Bandelt HJ, Kivisild T, Parik J *et al.* (2006) Lab-specific mutation processes. In: *Human Mitochondrial DNA and the Evolution of Homo Sapiens* (eds Bandelt H, Macaulay V, Richards M), pp. 117–146. Springer, Berlin, Heidelberg.

Bandelt HJ, Kloss-Brandstätter A, Richards MB, Yao YG, Logan I (2014) The case for the continuing use of the revised Cambridge Reference Sequence (rCRS) and the standardization of notation in human mitochondrial DNA studies. *Journal of Human Genetics*, **59**, 66–77.

Bonfiglio S, Achilli A, Olivieri A *et al.* (2010) The enigmatic origin of bovine mtDNA haplogroup R: sporadic interbreeding or an independent event of *Bos primigenius* domestication in Italy? *PLoS ONE*, **5**, e15760.

Bonfiglio S, Ginja C, De Gaetano A *et al.* (2012) Origin and spread of *Bos taurus*: new clues from mitochondrial genomes belonging to haplogroup T1. *PLoS ONE*, **7**, e38601.

Doro MG, Piras D, Leoni GG *et al.* (2014) Phylogeny and patterns of diversity of goat mtDNA haplogroup A revealed by resequencing complete mitogenomes. *PLoS ONE*, **9**, e95969.

Hassanin A, Bonillo C, Nguyen BX, Cruaud C (2010) Comparisons between mitochondrial genomes of domestic goat (*Capra hircus*) reveal the presence of numts and multiple sequencing errors. *Mitochondrial DNA*, **21**, 68–76.

Kong QP, Salas A, Sun C *et al.* (2008) Distilling artificial recombinants from large sets of complete mtDNA genomes. *PLoS ONE*, **3**, e3016.

Lancioni H, Di Lorenzo P, Ceccobelli S *et al.* (2013) Phylogenetic relationships of three Italian merino-derived sheep breeds evaluated through a complete mitogenome analysis. *PLoS ONE*, **8**, e73712.

Liu J, Wang LD, Sun YB *et al.* (2012) Deciphering the signature of selective constraints on cancerous mitochondrial genome. *Molecular Biology and Evolution*, **29**, 1255–1261.

Miao YW, Peng MS, Wu GS *et al.* (2013) Chicken domestication: an updated perspective based on mitochondrial genomes. *Heredity*, **110**, 277–282.

van Oven M, Kayser M (2009) Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation. *Human Mutation*, **30**, E386–E394.

Pang JF, Kluetsch C, Zou XJ *et al.* (2009) mtDNA data indicate a single origin for dogs south of Yangtze River, less than 16,300 years ago, from numerous wolves. *Molecular Biology and Evolution*, **26**, 2849–2864.

Salas A, Carracedo A, Macaulay V, Richards M, Bandelt HJ (2005a) A practical guide to mitochondrial DNA error prevention in clinical, forensic, and population genetics. *Biochemical and Biophysical Research Communications*, **335**, 891–899.

Salas A, Yao YG, Macaulay V *et al.* (2005b) A critical reassessment of the role of mitochondria in tumorigenesis. *PLoS Medicine*, **2**, e296.

Salas A, Coble M, Desmyter S *et al.* (2012) A cautionary note on switching mitochondrial DNA reference sequences in forensic genetics. *Forensic Science International Genetics*, **6**, e182–e184.

Wang Z, Shen X, Liu B *et al.* (2010) Phylogeographical analyses of domestic and wild yaks based on mitochondrial DNA: new data and reappraisal. *Journal of Biogeography*, **37**, 2332–2344.

Wang G-D, Xie H-B, Peng M-S, Irwin D, Zhang Y-P (2014) Domestication genomics: evidence from animals. *Annual Review of Animal Biosciences*, **2**, 65–84.

Wu GS, Yao YG, Qu KX *et al.* (2007) Population phylogenomic analysis of mitochondrial DNA in wild boars and domestic pigs revealed multiple domestication events in East Asia. *Genome Biology*, **8**, R245.

Yao YG, Kong QP, Salas A, Bandelt HJ (2008) Pseudomitochondrial genome haunts disease studies. *Journal of Medical Genetics*, **45**, 769–772.

Yao YG, Salas A, Logan I, Bandelt HJ (2009) mtDNA data mining in GenBank needs surveying. *American Journal of Human Genetics*, **85**, 929–933; author reply 933.

Yu G, Xiang H, Wang J, Zhao X (2013) The phylogenetic status of typical Chinese native pigs: analyzed by Asian and European pig mitochondrial genome sequences. *Journal of Animal Science Biotechnology*, **4**, 9.

M.-S.P. and Y.-P.Z. designed research, N.-N.S. conduct the experiment, N.-N.S., L.F. and M.-S.P. preformed the analysis, N.-N.S., Y.-G.Y., M.-S.P. and Y.-P.Z. wrote the manuscript.

doi: 10.1111/mec.12955

Data accessibility

The mtDNA sequence alignments were deposited in Dryad (doi:10.5061/dryad.cc5kn).

Supporting information

Additional supporting information may be found in the online version of this article.

Table S1 Access numbers of (near-)complete mtDNA genome sequences.

Table S2 Potential errors identified in eight domestic animals.

Table S3 mtDNA haplogroup tree of dog and grey wolf.

Table S4 mtDNA haplogroup tree of goat.

Table S5 mtDNA haplogroup tree of yak.

Table S6 Potential errors detected in next-generation sequencing data.