Vietnam Journal of Biotechnology **21**(4): 655-667, 2023

MITOCHONDRIAL DNA INSERTIONS INTO NUCLEAR GENOME (NUMTS) IN VIETNAM SHEEP: IMPLICATIONS AND CAVEATS FOR THE UTILITY OF THE MITOCHONDRIAL DNA MARKERS IN GENETIC DIVERSITY STUDIES IN SHEEP

Ngoc Luong Nguyen[1,✉], **Nhu Cuong Nguyen**[1,2], **Thi Thu Ha Huynh**[1]

[1]*Department of Biology, College of Sciences, Hue University, 77 Nguyen Hue Road, Phu Nhuan Ward, Hue City, Thua Thien Hue Province, Vietnam*
[2]*Chu Van An Secondary School, 36A Duong Van An Road, Xuan Phu Ward, Hue City, Thua Thien Hue Province, Vietnam*

✉To whom correspondence should be addressed. E-mail: luongnguyenbio@hueuni.edu.vn

Received: 02.11.2023
Accepted: 25.12.2023

SUMMARY

Nuclear insertions of mitochondrial DNA sequences (numts) can introduce errors in population genetic studies using mitochondrial DNA markers (mtDNA markers). Despite their prevalence in mammalian genomes and the potential risk of contamination with mtDNA markers, adequate precautions are often overlooked. When investigating the genetic diversity of Vietnam sheep, we found numt contamination in D-loop amplification. Consequentially, to continue, we conducted a comprehensive *in silico* survey of numts using the reference genomes from the NCBI database. We identified four so-called "mega-numts" that are not only highly similar in sequence to the mitogenome but also span a large part of it, thus posing a serious problem for amplifying authentic mtDNA markers. We demonstrated the existence of two numts in three randomly selected Vietnam sheep individuals. Based on the assumed mutation rate of the D-loop region, we proposed a sequence identity threshold to help distinguish authentic D-loop sequences from their corresponding numts when studying contemporary sheep populations. We hope this study can serve as a guideline for future research into Vietnam sheep using mtDNA markers.

Keywords: Contamination, D-loop, Dorper sheep breed, numt, Phan Rang sheep breed, Vietnam sheep

INTRODUCTION

Mitochondria, the powerhouse of eukaryotic cells, are theorized to have originated from a symbiotic relationship with an alpha-proteobacterium. These organelles possess their own distinct DNA, known as mitochondrial DNA (mtDNA) or mitogenome, which stands apart from the nuclear genome. Since the pioneering studies by Avise (1986) and Mortiz *et al.* (1987), mtDNA markers have risen in prominence for tracing matrilineal inheritance, relying on principles of clonal inheritance, presumed neutrality, and a consistent mutation rate (Galtier *et al*., 2009). Despite these

advantages, the indiscriminate application of mtDNA markers raises concerns, particularly due to contamination by nuclear-inserted mitochondrial sequences, (numts). A survey of mammalian genomes illustrates the prevalence of numts, ranging from a single insertion in *Caenorhabditis. elegans* to a multitude in the platypus, with phylogenetic studies dating insertions from as recent as 680 thousand years ago to as far as 97.5 million years ago (Calabrese *et al*., 2017).

Numts and recommendations for caution in the use of mtDNA markers have been documented across various species, including humans (Simone *et al*., 2011; Xue *et al*., 2023), goats (Hassanin *et al*., 2010; Ning *et al*., 2017), cattle (Liu, Zhao, 2007; Grau *et al*., 2020), and birds (Lucas *et al*., 2022; Baltazar-Soares *et al*., 2023). In the realm of sheep genetics, Féménia *et al*. (2021) conducted an *in silico* survey of numts using the reference genomes from NCBI GenBank, and Mustafa *et al.* (2018) highlighted the potential for mtDNA marker contamination by numts in genetic studies. Despite these issues, mtDNA markers remain a tool of choice for many geneticists and ecologists due to their cost-effectiveness compared to whole mitogenome or whole nuclear genome analyses (Galtier *et al*., 2009; Ibrahim *et al*., 2023). As for sheep, several previous studies have made use of the D-loop region to study the genetic diversity of sheep (Zhao *et al*., 2011; Zhao *et al*., 2013; Othman *et al*., 2015; Ibrahim *et al*., 2020; Kamalakkannan *et al*., 2021; Germot *et al*., 2022; Salim *et al*., 2023)

The Phan Rang sheep, often considered a native Vietnamese breed, has debated origins, with some local accounts suggesting introduction from India via Malaysia Kelantan region, while other sources point to importation from Europe during French colonial times (Le, Le, 2005; Doan *et al*., 2006). This breed, named after its initial rearing location in Phan Rang City, Ninh Thuan Province, had been traditionally confined to the regions for nearly a century. Recent efforts to expand their range have achieved mixed results (Bui, 2014; Ngo, 2014).

Amidst efforts by the local authorities to enhance the productivity and disease resistance of Phan Rang sheep through cross-breeding with newly imported breeds such as Australian Dorper or White Suffolk rams (Le, Le, 2005), there is a looming risk of diluting or even losing the unique genetic resources of the Phan Rang sheep. Our research investigated the genetic makeup of the Phan Rang sheep, particularly focusing on maternal lineages by employing the control region (D-loop). However, during our cloning attempts of the D-loop from the Phan Rang and Dorper sheep, we were confounded by numt contamination, leading us to conduct a comprehensive examination of numts in contemporary sheep breeds in general and Vietnamese sheep breeds in particular. This report presents our preliminary findings on numts within both the reference sheep genomes and the Phan Rang sheep and Dorper sheep bred in Vietnam. Our objective is to arm future researchers with the necessary insights to circumvent the challenges posed by numts in the genetic analysis of Vietnamese sheep breeds.

MATERIALS AND METHODS

Sample collection

The study used blood samples obtained from 57 ewes, of which 49 were from Ninh Thuan Province and 8 from Son Tay's Goat and Rabbit Research Center. Among these

53 ewes were Phan Rang sheep and 4 ewes were Dorper. The Son Tay's Goat and Rabbit Research Center has been housing Phan Rang sheep for decades. Approximately 2 mL of blood were taken from the jugular veins of sheep by the certified local veterinarians. The blood samples were preserved in EDTA anticoagulant tubes at 4 ^{o}C to 8 ^{o}C until use. For long-term storage, the blood samples were stored at –80 ^{o}C.

DNA extraction and PCR conditions

Total DNA was extracted using the FavoPrepTM Blood/Culture Cell Genomic DNA Extraction Mini Kit (Farvogen, Cat No. FABGK 001-2). DNA concentrations were determined at 260/280 nm with the spectrophotometer and 0.8% agarose gel electrophoresis.

All PCR reactions in this study employed proofread Taq polymerase. In particular, all D-loop samples were amplified with Phusion High-Fidelity Taq Polymerase (Thermofisher Scientific, Cat No. F350S). For fragments over 3 kb, Plantinum^{TM} SuperFi^{TM} DNA Polymerase (Thermofisher Scientific Cat No. 12351010) was used. To minimize numts contamination, for D-loop amplification approximately 30–50 ng of total DNA was used as a template, while for numt amplification 100–150 ng of DNA was used as template.

The settings for D-loop amplification are as follows: Denaturing at 98 ^{o}C for 2 min followed by 30 cycles of 98 ^{o}C for 30 sec, 60 ^{o}C for 30 sec and 72 ^{o}C for 30 sec. It ended with 10 mins at 72 ^{o}C followed by 10 mins at 4 ^{o}C. The settings for larger fragments (>7 kb) from mtDNA are as follows: Denaturing 98 ^{o}C for 2 mins followed by 10 cycles of 98 ^{o}C for 30 sec, 62 ^{o}C for 30 sec and 72 ^{o}C for 3 mins (15–30 sec per kb), followed by 25 cycles of 98 ^{o}C for 30 sec, 62 ^{o}C for 30 sec and 72 ^{o}C for 5 mins. The reaction ended with 72 ^{o}C for 10 mins followed by 4 ^{o}C for 10 mins. For numts amplification (~ 3 kb), the same setting for mitogenome fragment amplification was utilized, but the amount of template was increased threefold.

The primer pairs used for the amplifications above are provided in table 1. All PCR products were analyzed on a 1% agarose gel and purified using the Favorgen Gel/PCR Purification Kit (Cat No. FAGCK 001-1).

Table 1. Sequences of the primers used in this study.

Primer name	Primer sequences (5' → 3')
D-loop F	CCAGAGAAGGAGAACAACCAA
D-loop R	GCATTTTCAGTGCCTTGCTT
F1-F	TCACCATTTCGGTTTCGAAGCC
F1-R	TGTTACGACTTGTCTCCTCTCG
NumtsX-F	TGGTGTGCTAGGGAATCTTGAC
NumtsX-R	CCGGTAGTACTCTGGCGAATAA
Numts6-F1	GAGTTAAAGCTATAATCCTTACAGTCC
Numts6-R1	CCGATTAGGTTGATTGATGGG

Cloning, sequencing and analysis

The PCR products from the amplification of the D-loop region were submitted for direct sequencing to Asia FirstBase using amplification primers. In cases where sequencing results were of suboptimal quality, indicated by the presence of double peaks or insufficient length (<1000 bp), the PCR products were then cloned into pGEMT Easy vector for subsequent resequencing. The chromatogram files were visually examined, and low-quality regions and/or vector regions were trimmed before the forward and reverse sequencing results were joined into contigs using Snapgene version 7. From Snapgene, sequences were extracted into a fasta format for GenBank submission. All the D-loop sequences and contaminated numts sequences were submitted to GenBank through Bankit for accession numbers. The sequences were blasted against the GenBank database to ascertain the identity of D-loop, which is limited to domestic sheep sequences. Discontiguous BLAST was used to enhance the sensitivity and specificity of the search. The definitive proof of D-loop authenticity is established by a sequence identity of 98% or higher, coupled with 100% coverage when compared to any sequence in the GenBank database.

The D-loop sequences of the Phan Rang sheep were aligned for the pairwise distances by using MEGA 11 (Tamura *et al*., 2021) using the MUSCLE algorithm. The mega-format file was exported and used as an input for pairwise distance matrix construction using MEGA 11. Subsequently, the matrix was converted into a histogram of sequence identity percentages by Excel. The same procedure was applied when determining the maximum and minimum sequence identity percentages among domestic sheep and wild sheep.

***In silico* analysis of numts and D-loop**

We conducted an *in silico* analysis to identify nuclear mitochondrial DNA segments (numts) within the sheep nuclear genome. Utilizing the reference sheep mitochondrial genome (accession number NC_00194.1) as our query sequence, we searched against the sheep reference nuclear genome (GCF_016772045.1). To accommodate the circular structure of the mitogenome vs. linear structure of numts, and to ensure identification of the complete numts sequences, we systematically shifted the starting base of the mitogenome by intervals of 1000 bases. Discontiguous BLAST was employed to enhance sensitivity. From the BLAST results, we selected so-called mega-numts, which are long in sequence and share a high sequence identity with the mitogenome.

To establish a sequence identity threshold for distinguishing authentic D-loop sequences from their numts counterparts, we used the D-loop region from the reference mitogenome as a query in a BLAST search against the sheep nucleotide sequence database, in another word NCBI GenBank limited to domestic sheep sequences. The BLAST results were sorted for the nucleotide sequence identity, categorized into bins, and visually presented in a plot generated using Microsoft Excel. To estimate the minimum percentage of sequence identity between D-loop regions from two distinct sheep breeds, we made several assumptions. First, since the rate of mutation of the D-loop region in sheep is not available, we extrapolated the rate of mutation of the D-loop region from cattle

(Mona *et al*., 2010). Second, we adopted the theory that sheep were domesticated approximately 12,000 years ago (Lv *et al*., 2015; Deng *et al*., 2020). We then calculated the expected number of mutations accumulated over the domestication period using the mutation rate per nucleotide and the total nucleotide count of the complete D-loop region. Hence, the percentage sequence identity (PSI) is calculated as follows:

$$PSI = 100\% \frac{(\text{Total D-loop nucleotide count} - 2 \times \text{total No of mutations})}{\text{Total D-loop nucleotide count}}$$

Where the total number of mutations can be calculated from the formula

Total number of mutations = (mutation rate/site/year) x number of years x total D-loop nucleotide count.

Two means that for any two contemporary sheep breeds evolving independently from the common ancestor, the maximum number of mutations will be the sum of the total number of mutations from each breed.

Therefore, given any two contemporary sheep breeds, the maximum number of nucleotide differences in their D-loop regions is: 2 x [(mutation rate/site/year) x 12,000 years x 1180].

Demonstration of numts presence in the Phan Rang sheep nuclear genomes

To establish the proof of numts presence in Phan Rang sheep, we designed two pairs of primers targeting regions on chromosome 6 (Chr. 6) and chromosome X (Chr. X). These regions span the flanking region outside the corresponding numts at the 5' end and extend into the numts at the 3' end. PrimerBLAST was employed to design these primers. Three sheep were randomly selected for PCR. The results were analyzed on a 1% agarose gel.

RESULTS AND DISCUSSION

Sample collection and total DNA extraction

Total DNA (100 µ L for each) was successfully extracted from the blood of 57 ewes. A single band over 10 kb was visualized on the 0.8% agarose gel with little or no smearing (Figure 1A). The concentration of total DNA ranged from 35 to 50 ng/µL with 260/280 ratios within 1.9–2.0. The DNA samples were stored at –80 ºC for subsequent experiments.

The D-loop sequences from the Phan Rang sheep

All D-loop amplification attempts consistently produced bands of approximately 1.2 kb (Figure 1B). The final sequences were obtained after sequencing of the direct PCR products or clones and deposited in the NCBI GenBank database, with accession numbers: OR683520 to OR683588. Analysis of the 57 obtained sequences revealed three with notably low sequence identity to any D-loop sequences in GenBank (below 90%). BLAST searches against the reference nuclear genome revealed that two sequences (OR683589 and OR683590, belonging to the Phan Rang breed) shared 99.6% sequence identity with

the numts on Chr. X of the Rambouillet nuclear reference genome. The third sequence (OR683591, from the Dorper breed) showed only 91% identity with a numts on Chr. 2, suggesting numts contamination. To circumvent this, the contaminated D-loop sequences were reamplified using an alternative strategy that entailed the initial amplification of a larger fragment of the mitogenome encompassing the D-loop region but not fully overlapping with any known numts. These were then used as templates for subsequent D-loop amplification. The corrected sequences have been deposited in GenBank with accession numbers OR921192 to OR921194.

The presence of numts in sheep has been previously documented (Féménia *et al*., 2021). However, our findings highlight a significant knowledge gap, as only numts from the reference genome are readily available, and the numts from the Dorper breed exhibit just 91% identity with the reference numts. This underscores the need for further research into numts across diverse sheep breeds, as the consequences of misidentification or overlooking numts contamination in mtDNA-based genetic studies can lead to substantial inaccuracies in ovine populations.

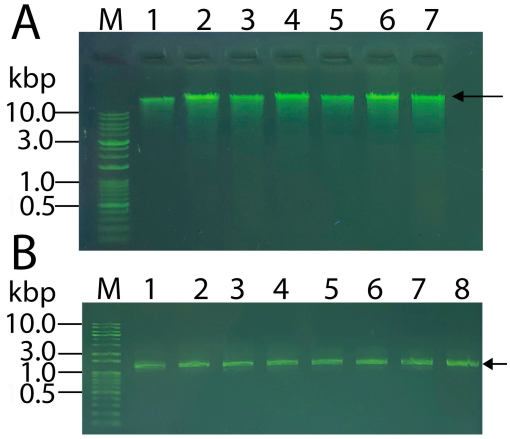


Figure 1. Representative results of total genomic DNA extraction and D-loop PCR from Vietnamese sheep. A. Total genomic DNA extracted from sheep whole blood; B. D-loop PCR results. The arrows indicate the expected positions.

In silico analysis of numts from the reference genomes

Following the work of Féménia *et al*., (2021), we attempted to thoroughly survey the numts in sheep using the reference mitogenome and nuclear genome. We particularly paid close attention to numts that share high sequence similarity with the mitogenome, particularly in regions corresponding with popular mtDNA markers such as D-loops, CtyB, and COX1. Table 2 shows some numts that have been identified as posing a risk of contaminating mtDNA markers and figure 2 shows these numts mapped onto the mitogenome. We are interested in so-called mega-numts, which include numts on Chr. 6, Chr. X, and Chr. 12 and span more than one-third of the mitogenome and share a very high sequence identity with them. These numts also contain regions that are highly similar to D-loop, CytB, and COX1, which can potentially lead to misidentification. Besides these mega-numts,

which may hinder the amplification of full mitogenome, there are small numts scattering across the genomes, posing a contamination risk to studies involving amplifying a particular mtDNA marker. The OR683591 sequence is one example of such small numts.

Table 2. Some mega-numts, their positions on the chromosome, their sequence identities compared to the mitogenome, and their sequence identities compared to the mtDNA markers.

Positions on chromosome	Sequence identity with the mitogenome	Sequence identity with common mtDNA markers
Chr. X (59766804…59780263)	98%	97.99% (COX)
Chr. 6 (27830487… 27841033)	99%	98.68% (CytB), 97.3% (D-loop)
Chr. 2 (55748565… 55758677)	93%	92.1% (COX)
Chr. 12 (27346074… 27353266)	97%	97.02% (COX)
Chr. 4 (77469321… 77477352)	92%	91.59% (COX)

Figure 2. Mega-numts mapped to the mitogenome. The positions of D-loop primers and the primers used to amplify the fragment containing the D-loop without co-amplifying numts were shown (Drawn using Snapgene version 7).

Confirmation of numts presence in Phan Rang sheep

To obtain further evidence of numts presence in Phan Rang sheep, we designed two pairs of primers that exclusively amplify the regions on Chr. 6 and Chr. X without co-amplifying the mitogenome. The primer sequences can be found in table 1. We tested the primer pairs on three randomly selected sheep. The results are shown in figure 3. Although we have not sequenced the PCR products, the fact that PCR results turn out positive partly confirms the presence of numts in the Phan Rang sheep genomes.

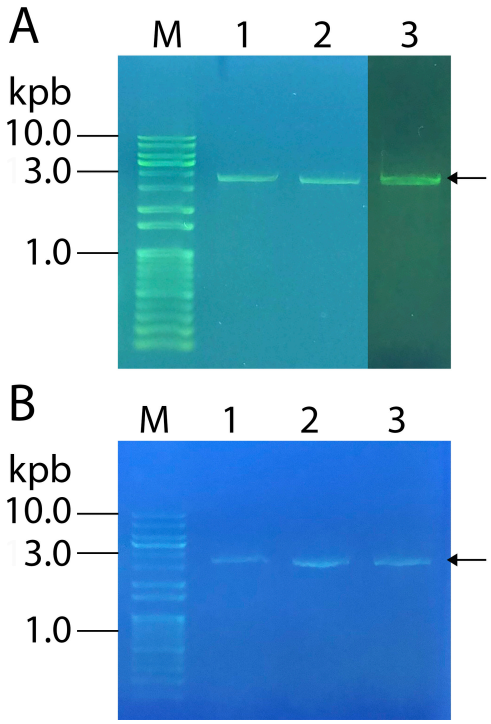


Figure 3. Demonstration of numts presence in three randomly selected Phan Rang sheep using PCR. A Partial numts from Chr. X and B Partial numts from Chr. 6.

Establishing guidelines for authentication of D-loop amplification

Given the widespread use of mtDNA markers in sheep genetic studies (Zhao *et al*., 2011; Zhao *et al*., 2013; Othman *et al*., 2015; Ibrahim *et al*., 2020; Kamalakkannan *et al*., 2021; Germot *et al*., 2022; Salim *et al*., 2023), caution against numts contamination is essential. From a literature review, the suggestions for use due to this caution of numts contamination are not always exercised. For instance, Lv *et al*., (2015) used 21 primer pairs to amplify the sheep mitogenome, with many potentially amplifying numts. Likewise, Niu *et al*. (2017) utilized 15 primer pairs for the Tibetan sheep mitogenome without addressing numts contamination, risking inaccurate variance calling.

To investigate whether numts might have been reported as D-loop in NCBI GenBank, we used the reference mitogenome as the query to search for homologous sequences in the nucleotide collection, i.e., GenBank.

When ranking these sequences according to their sequence identity, we observed a sudden drop in the number of sequences with identities lower than 98% and lower than 95% (see Figure 4).

Given that sheep have only been domesticated for about 12,000 years at maximum, we can establish another line of evidence to authenticate the D-loop. Given that the length of the D-loop is 1180 base pairs, and the mutation rate of approximately $7.6x10^{-7}$ per site per year, extrapolated from cattle (Mona *et al*., 2010), we deduced that any two given contemporary sheep breeds would have a total of 22 nucleotide differences at most in their D-loop regions. This amounts to a minimum identity of approximately 98%.

The plot derived from the pairwise distance matrix of the Phan Rang sheep showed a somewhat similar pattern to the plot from the BLAST search (see Figure 4). Most Phan Rang sheep share a percentage sequence identity above 96%. Although Phan Rang sheep were imported into Vietnam just over a century ago, they came from many sources; therefore, their mtDNA should reflect the diversity of their origins.

Substantial evidence has been gathered to support a consensus theory about sheep domestication and migration. In terms of mtDNA, sheep can be thought of as consisting of two major haplogroups, haplogroup A and haplogroup B, with haplogroup A dominating Europe and Africa, while haplogroup B dominating Asia. Haplogroup B migrated from the region around present-day Turkey to Mongolia then split into two subgroups, in which one migrated to India and the other to China (Lv *et al*., 2015; Deng *et al*., 2020; Macho vá *et al*., 2022). Therefore, we can find another line of evidence to establish the authenticity of the D-loop sequences. By determining the maximum and minimum percentage sequence identity among domestic sheep breeds from different geographical regions the values were corresponded to two major sheep haplogroups. Table 3 shows the pairwise distances among domestic sheep breeds and among domestic sheep breeds with their wild cousins. The highest sequence identity is between the Bashbay breed and the Merino breed (99.4%) while the lowest is between the Yecheng breed and the wild sheep (93.9%). Sheep belonging to the same haplogroup share a sequence identity higher than 99%, while sheep between haplogroups share a sequence identity higher than 96%.

Given the extensive sequencing of D-loop regions across various contemporary sheep breeds, whose sequences are readily accessible in GenBank, it is highly unlikely that a D-loop sequence from an unknown domestic sheep breed would not exhibit high similarity to existing sequences in the database. Taken together, we suggest that when a BLAST search of a putative D-loop sequence yields hits with a percentage sequence identity lower than 98%, it's recommended to suspect numts contamination.

One major limitation of this study is that we have not addressed the problem of mega-numts, which may interfere with mitogenome enrichment and sequencing. Mega-numts span more than one-third of the mitogenome and can easily be mistaken for the latter if due care is not taken. Whole mitogenome sequencing therefore should take into account numts as well as mega-numts to eliminate false variance callings.

Table 3. Pairwise percentage sequence identity among contemporary domestic sheep breeds and their wild cousin. The highest and lowest sequence identities are in bold. The accession number for the mitogenome of the breeds can be found below.

	Bashbay	**Yecheng**	**Merino**	**Turkey**	**Oxford**	**Swiniarka**	**Sunite**
Bashbay							
Yecheng	99.4%						
Merino	**99.8%**	99.4%					
Turkey	96.4%	96.6%	96.6%				
Oxford	96.2%	96.4%	96.4%	99.6%			
Swiniarka	96.4%	96.4%	96.6%	99.8%	99.6%		
Sunite	99.7%	99.3%	96.6%	99.6%	96.4%	96.6%	
Wild_sheep	94.3%	**93.9%**	94.55	94.2%	94.0%	94.2%	94.7%

Note: Accession number for the breeds: Bashbay (KF938330), Yecheng (KF938338), Merino Australia (HM236174), Karakas Turkey (HM236176), Oxford_down (KF938359), Swiniarka (KF938349), Sunite (KF938317), Wild sheep/*Ovis vignei* (HM236186). The highest and lowest percentage sequence identities among the sheep being compared are bolded.

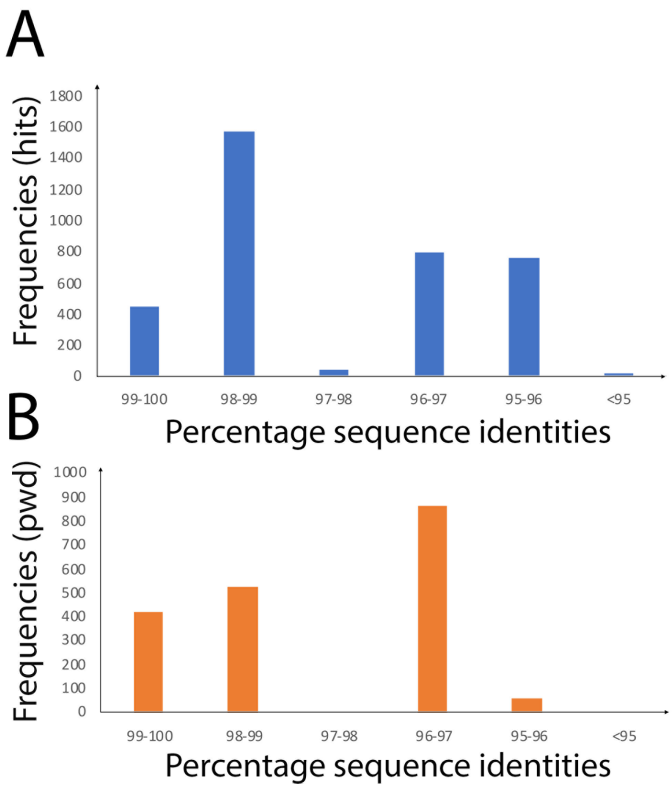


Figure 4. Plots of frequencies of sequence identity for the D-loop region of sheep. A. Frequencies of D-loop by percentage sequence identities between the D-loop from the reference mitogenome and other D-loop sequences available in GenBank; B. Frequencies of D-loop by percentage sequence identity among Phan Rang sheep in this study. (pwd) = pairwise distances)

CONCLUSION

In this study, prompted by the incident of numts contamination in our Phan Rang sheep genetic study using the D-loop as the marker, we conducted a comprehensive *in silico* survey of numts in the sheep nuclear genome. Four mega-numts were identified on Chr. X, Chr. 6, Chr. 2, and Chr. 12. From the literature survey, it can be seen that the issue of numts contamination in sheep has not been taken with sufficient precaution. The fact is that one of our numts is slightly different from the numts of the reference genome indicates potential variance of numts sequences among sheep breeds. The presence of two numts in three randomly selected Phan Rang sheep was demonstrated. Our method can be extended to other popular mtDNA markers, such as Cytb and COX1.

Acknowledgement: *This study is supported by the Grant "Sequencing of Phan Rang sheep mitogenome,", code CT-2021-01-DDH-01, which belongs to the Program "Application of biotechnology to exploit and develop some plant and animal breeds in Central and Highland Vietnam" from the Ministry of Education.*

REFERENCES

Avise JC (1986) Mitochondrial DNA and the evolutionary genetics of higher animals. *Philos Trans R Soc Lond B Biol Sci* 312: 325–342.

Baltazar-Soares M, Karell P, Wright D, Nilsson JA, Brommer JE (2023) Bringing to light nuclear-mitochondrial insertions in the genomes of nocturnal predatory birds. *Mol Phylogenet Evol* 181: 107722.

Bui VL (2014) Evaluation of growth performance of Phan Rang raised in Thua Thien Hue Province [PhD dissertation] [Hue, Thua Thien Hue], Hue University. 174 pages.

Calabrese FM, Balacco DL, Preste R, Diroma MA, Forino R, Ventura M, Attimonelli M (2017) NumtS conlonization of mammalian genomes. *Sci Rep* 7: 16357.

Deng J, Xie XL, Wang DF, Zhao C, Lv FH, Li X, Yang J, Yu JL, Shen M, Gao L *et al.* (2020) Paternal origins and migratory episodes of domestic sheep. *Curr Biol* 30: 4085–4095.e6.

Doan DV, Vuong NL, Ho QA (2006) An evaluation of Phan Rang sheep conformation. *Vietnam J Animal Sci Technol* 10: 11–13.

Féménia M, Charles M, Boulling A, Rocha D (2021) Identification and characterization of mitochondrial sequences integrated into the ovine nuclear genome. *Anim Genet* 52: 556–59.

Galtier N, Nabholz B, Glémin S, Hurst GDD (2009) Mitochondrial DNA as a marker of molecular diversity: a reappraisal. *Mol Ecol* 18: 4541–4550.

Grau ET, Charles M, Féménia M, Rebours E, Vaiman A, Rocha D *et al.* (2020) Survey of mitochondrial sequences integrated into the bovine nuclear genome. *Sci Rep* 10: 2077.

Germot A, Khodary MG, Othman OE, Petit D (2020) Shedding Light on the Origin of Egyptian Sheep Breeds by Evolutionary Comparison of Mitochondrial D-Loop. *Animals* (Basel): 2738.

Hassanin A, Bonillo C, Bui XN, Cruaud C (2010) Comparisons between mitochondrial genomes of domestic goat (*Capra hircus*) reveal the presence of numts and multiple sequencing errors. *Mitochondrial DNA* 2: 68–76.

Ibrahim A, Budisatria IGS, Widayanti R, Artama WT (2020) The genetic profiles and maternal origin of local sheep breeds on Java Island (Indonesia) based on complete mitochondrial DNA D-loop sequences. *Vet World* 13: 2625–2634.

Ibrahim A, Baliarti E, Budisatria IGS, Armata WT, Widayanti R, Maharani D, Tavarres L, Margawati ET (2023) Genetic diversity and relationship among Indonesian local sheep breeds on Java Island based

on the mitochondrial cytochrome b gene sequences. *J Genen Eng Biotechnol* 21: 34.

Kamalakkannan R, Kumar S, Bhavana K, Prabhu VR, Machado CB, Singha HS, Sureshgopi D, Vijay V, Nagarajan M (2021) Evidence for independent domestication of sheep mtDNA lineage A in India and introduction of lineage B through Arabian sea route. *Sci Rep* 11: 19733.

Le MC, Le DD (2005) Chăn nuôi cừu (Sheep raising techniques). *NXB Nông nghiệp* (Agriculture Publishing House), pp. 12–14.

Liu Y and Zhao Y (2007) Distribution of nuclear mitochondrial DNA in cattle nuclear genome. *J Anim Breed Genet* 124: 264–268.

Lucas T, Vincent B, Eric P (2022) Translocation of mitochondrial DNA into the nuclear genome blurs phylogeographic and conservation genetic studies in seabirds. *R Soc Open Sci* 9: 211888.

Lv FH, Peng WF, Yang J, Zhao YX, Li WR, Liu MJ, Ma YH, Zhao QJ, Yang GL, Wang F *et al.* (2015) Mitogenomic Meta-analysis identifies two phases of migration in the history of Eastern Eurasian Sheep. *Mol Bio Evol* 32: 2515–2533.

Machová K, Málková A, Vostrý L (2022) Sheep Post-Domestication Expansion in the Context of Mitochondrial and Y Chromosome Haplogroups and Haplotypes. *Genes* 13: 613.

Mortiz C, Dowling TE, Brown WM (1987) Evolution of animal mitochondrial DNA: relevance for population biology and systematics. *Ann Rev Ecol Syst* 18: 269–292.

Mustafa SI, Schwarzacher T, Heslop-Harrison JS (2018) Complete mitogenomes from Kurdistani sheep: abundant centromeric nuclear copies representing diverse ancestors. *Mitochondrial DNA A* 29: 1180–1193.

Mona S, Catalano G, Lari M, Larson G, Boscato P, Casoli A, Sineo L, Di Patti C, Pecchioli E, Caramelli D, Bertorelle G (2010) Population dynamic of the extinct European aurochs: genetic evidence of a north-south differentiation pattern and no evidence sof post-glacial expansion. *BMC Evol Biol* 26: 83.

Ngo TV (2014) Investigating growth, reproduction and carcass performance of the Phan Rang sheep and some solutions to improve its meat productivity [PhD dissertation] [Ha Noi], National Institute of Animal Science. 170 pages.

Ning FY, Fu J, Du ZH (2017) Mitochondrial DNA insertions in the nuclear *Capra hircus* genome. *Genet Mol Res* 16: 1–8.

Niu L, Chen X, Xiao P, Zhao Q, Zhou J, Hu J, Sun H, Guo J, Li L, Wang L, Zhang H, Zhong T (2017) Detecting signatures of selection within the Tibetan sheep mitocondrial genome. *Mitochondrial DNA A DNA Mapp Seq Anal* 28: 801–809.

Othman OE, Pariset L, Balabel EA, Marioti M (2015) Genetic characterization of Egyptian and Italian sheep breeds using mitochondrial DNA. *J Genet Eng Biotechnol* 13: 79–86.

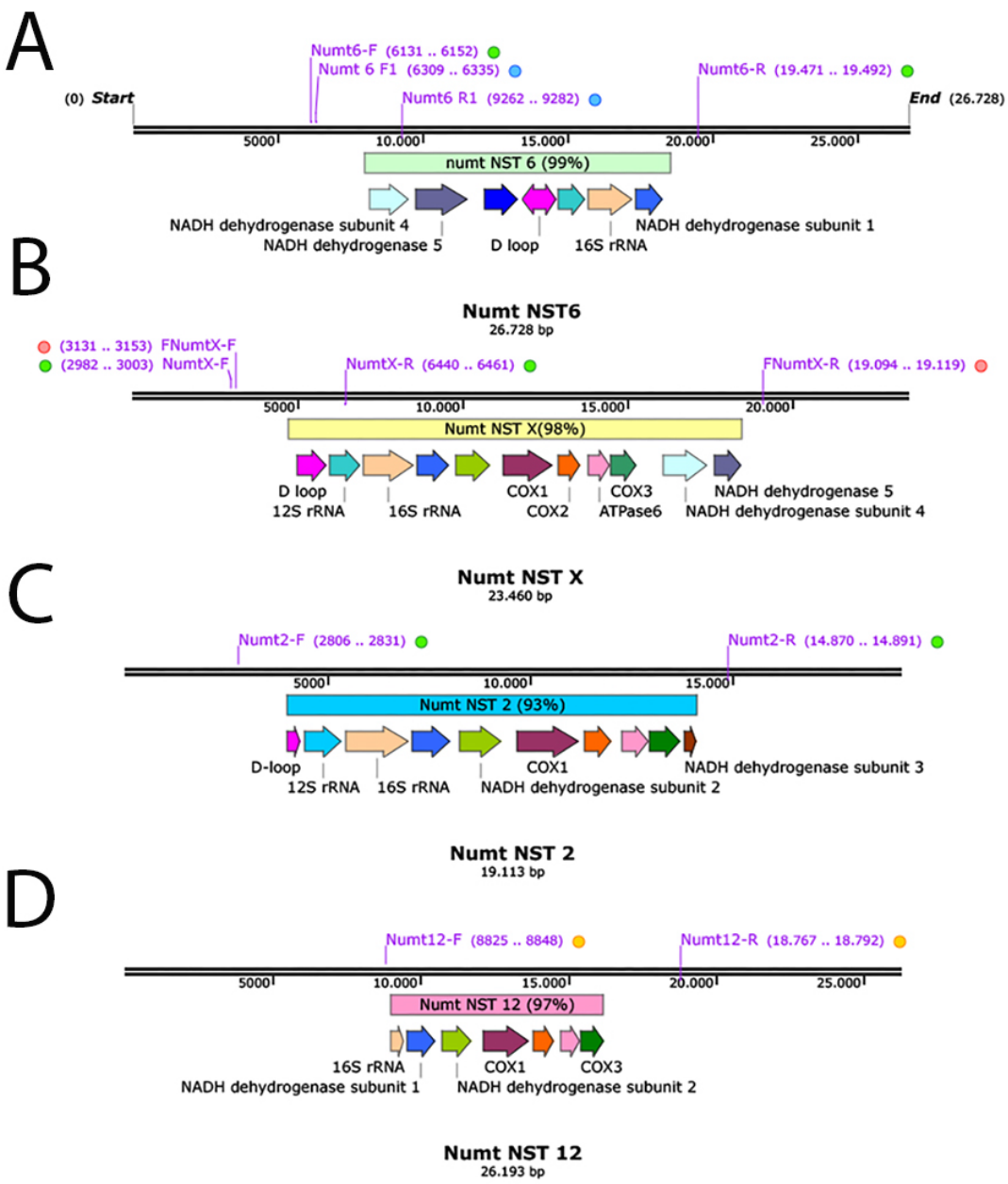
Salim B, Alasmari S, Mohamed NS, Ahmed MA, Nakao R, Hanotte O (2023) Genetic variation and demographic history of Sudan desert sheep reveal two diversified lineages. *BMC Genomics* 24: 118.

Simone D, Calabrese FM, Lang M, Gasparre G, Attimonelli M (2011) The reference human nuclear mitochondrial sequences compilation validated and implemented on the UCSC genome browser. *BMC Genomics* 12: 517.

Xue LY, Moreira JD, Smith KK, Fetterman JL (2023) The Mighty NUMT: Mitochondrial DNA Flexing its code in the nuclear genome. *Biomolecules* 13: 753.

Zhao E, Yu Q, Zhang N, Kong D, Zhao Y (2013) Mitochondrial DNA diversity and the origin of Chinese indigenous sheep. *Trop Anim Health Prod* 45: 1715–1722.

Zhao Y, Zhao E, Zhang N, Duan C (2011) Mitochondrial DNA diversity, origin, and phylogenic relationships of three Chinese large-fat-tailed sheep breeds. *Trop Anim Health Prod* 43: 1405–1410.



Supplementary Figure 1. Maps of mega-numts found using the reference mitogenome as a query searched against the reference nuclear genome. The positions of primers to amplify full sequences of the mega-numts as well as the positions of the primers used to demonstrate the presence of numts in Phan Rang sheep were shown. A. numts from Chr. 6; B. numts from Chr. X; C. numts from Chr. 2; D. numts from Chr. 12.