Identification and characterisation of mitochondrial sequences integrated into the ovine nuclear genome

M. Féménia, M. Charles, A. Boulling, D. Rocha

ANIMAL GENETICS Immunogenetics, Molecular Genetics and Functional Genomics

SHORT COMMUNICATION

doi: 10.1111/age.13096

Identification and characterisation of mitochondrial sequences integrated into the ovine nuclear genome

M. Féménia*, M. Charles*,†, A. Boulling* and D. Rocha*

*INRAE, AgroParisTech, GABI, Université Paris-Saclay, Jouy-en-Josas F-78350, France. †SIGENAE, INRAE, Jouy-en-Josas F-78350, France.

Summary

Mitochondrial DNA sequences are frequently transferred into the nuclear genome, generating nuclear mitochondrial DNA sequences (NUMTs). Here, we analysed, for the first time, NUMTs in the ovine genome. We obtained 760 alignment matches covering 513.8 kbp of the sheep nuclear genome. After a merging step, we identified 390 NUMT regions with a total length of ~720 kbp, representing 0.02% of the nuclear genome. We discovered copies of all mitochondrial regions and found that most NUMT regions are intergenic or intronic. Ovine NUMTs are mostly not transcribed. However, we identified within some of the NUMTs, potential new genes encoding nuclear *humanin* isoforms. To rule out the possibility that the identified NUMTs could be artifacts of the *Oar Rambouillet v1.0* genome assembly, we validated experimentally nine NUMT regions by PCR amplification. As we found several NUMT regions showing high similarity to the mitochondrial genome that potentially could pose a risk to ovine DNA mitochondrial studies, special care must be taken for the selection of primers for PCR amplification of mitochondrial DNA sequences.

Keywords sheep/pseudogene, mitochondria, genome

In eukaryotes, mitochondrial DNA sequences are frequently transferred into the nuclear genome, generating nuclear mitochondrial DNA sequences (NUMTs). This transfer has strongly influenced the evolution and function of eukaryotic genomes (Hazkani-Covo *et al.* 2010). Because of their homology, NUMTs can compromise mitochondrial DNA studies if they are not taken into account (Yao *et al.* 2008). It is therefore important to identify these nuclear sequences of mitochondrial origin. NUMTs have been documented in many mammals (Calabrese *et al.* 2017), including in cattle (Tramontin Grau *et al.* 2020) and goat (Ning *et al.* 2017), but not yet in sheep (*Ovis aries*). We present here the first study of NUMTs located into the ovine nuclear genome.

The mitochondrial reference sequence and the latest *Oar Rambouillet v1.0* reference genome sequence were retrieved from Genbank (assembly accessions: NC_001941.1 and GCA_002742125.1 respectively, 25 January 2021) for all analyses. Standard linear alignment tools such as BLASTN (Altschul *et al.* 1990) are relatively sensitive to the exact place where the genomic sequence begins. To handle the fact that the mitochondrial genome is circular, we perform alignments onto the ovine nuclear genome using two different linearised sequences of the ovine mitochondrial genome: (1) a standard linearisation starting at position 1 and ending at position 16 616; and (2) a shifted linearisation starting arbitrarily at position 1051 and ending at position 1050. This way, nuclear genomic sequences overlapping at the same time with both ends of the standard linear mitochondrial genome sequence could be identified. These two linearised mitogenome sequences were aligned to the sheep genome sequence using the program BLASTN following the procedure described in Calabrese *et al.* (2017). However, an *e*-value threshold of 10^{-4}, comparable to other studies of NUMTs (Hazkani-Covo *et al.* 2010) was applied. Alignment matches were detected in all chromosomes and in several unplaced scaffolds. However, as the nuclear genomic sequence was obtained from a female animal, the Y chromosome sequence is missing from the current sheep reference genome sequence (Fig. S1). We identified 760 NUMTs (alignment matches), with similarity between nuclear and mitochondrial sequences ranging between 63.7% and 99.9%. The total alignment length was 513.8 kbp, with alignment matches ranging from 36 to 14 381 bp. About 21% of the alignment matches between mitochondrial and nuclear sequences show sequence similarity higher than 85%. For example, NUMT S12.297, with a total length of 7193 bp, was detected in one contiguous alignment match, including only one gapped position and showing similarity of 96.7% to the mitochondrial sequence. It spans, without rearrangements, from position 2190 to

Address for correspondence

D. Rocha, INRAE, AgroParisTech, GABI, Université Paris-Saclay, Jouy-en-Josas F-78350, France.
E-mail: dominique.rocha@inrae.fr

Accepted for publication 08 May 2021

position 9385 of the mitogenome (Fig. 1). This dot plot graph was exported from the blast2seq NCBI website. Mitochondrial regions present in each NUMT were defined based on the positions of the alignment matches in the mitogenome. All mitochondrial genes had alignment matches with NUMTs, but distinct numbers of copies were detected (Fig. 2). The most occurring fragments of the mitogenome in the nuclear genome included *cytochrome oxidase* (*COX*) *I*, $tRNA^{Tyr}$, and the *D*-loop control region. The mechanism of mitochondrial DNA integration into the nuclear genome is not yet fully understood, it is therefore difficult to explain why some mitochondrial regions have more nuclear pseudogenes. The high number of nuclear sequences originating from mitochondrial *COXI*, $tRNA^{Tyr}$, and the *D*-loop region might be due to: (1) a higher number of direct integrations into the nuclear genome; (2) multiple duplications of these sequences post-integration; or (3) a combination of both. Interestingly Doynova *et al.* (2016) showed using chromosome conformation capture techniques to detect physical interactions between mitochondrial and nuclear DNA in human and mouse cells, that the *D*-loop region exhibited a higher tendency to interact with the nuclear genome. In addition, Rothfuss *et al.* (2010) have shown that the *D*-loop region is more susceptible to breakage than the rest of the mitochondrial genome raising the possibility that higher concentrations of *D*-loop fragments are released from complete mitochondrial genomes.

Nuclear copies of the mitogenome might be highly modified by insertions and deletions. Therefore, sequences resulting from one NUMT insertion event might be discovered as several alignment matches. NUMTs that were no more apart than 10 kbp on the nuclear genome were merged as one NUMT region. We identified after this merging step a total of 390 NUMT regions ranging from 36 to 35 131 bp (Table S1). NUMT regions with more than 5000 bp, although representing only around 11% of the total number of NUMTs regions, comprise around two thirds of the NUMT regions total length. By contrast, NUMT regions smaller than 300 bp comprise only about 6% of the total NUMT regions length, but represent 62% of the number of NUMT regions. Although the total alignment length reached ~514 kbp, the total length of the NUMT regions reached 717.9 kbp, representing 0.02% of the sheep nuclear genome. This represents an almost 40% increase of the total length detected after merging co-linear NUMTs. Interestingly we also found that 0.02% of the cattle genome included NUMT regions (Tramontin Grau *et al.* 2020). Some NUMT regions with sequence rearrangements, mostly insertions, are shown in Fig. S2. After excluding NUMT regions located on unmapped scaffolds, sequence comparisons among the remaining 251 NUMT regions revealed 4756 significant BLAST hits (identity ≥ 90%) between 198 different NUMTs, including 179 perfect matches (100% identity) between 61 different NUMT regions. This suggests that many of the identified NUMT regions are the result of duplications.

We downloaded from the ENSEMBL server the GTF file containing the genome annotation (release 103) to compare the location of genes and NUMT regions. NUMTs in sheep are mostly located in non-genic regions. We identified only 61 NUMT regions (~16%) overlapping with 60 genes (59 NUMT regions) and 2 pseudogenes (2 NUMT regions; Table S2). We found that 48 of the 59 NUMT regions overlapping with genes are located in introns. For example, NUMT region S7.212 is located within intron 5–6 of *NID2*. In addition, NUMT region S12.297 contains the upstream region of two neighbouring genes (*ENSOARG00020016617* and *ENSOARG00020017311*) while NUMT region S26.532 holds the downstream region of *NRG1* (*ENSOARG00020007228*). The remaining 10 NUMT regions carry the whole coding sequence of nine different genes. All these genes are encoding relatively small proteins with domains related to mitochondrial proteins. For example, *ENSOARG00020002046* is completely included within NUMT region S2.38. The 114 amino-acid protein encoded by this single-exon gene contains a ATPase domain. As these genes are only predicted *in silico*, experimental validation is required to confirm the existence of protein-coding genes.

Generally, NUMTs do not display transcriptional activity. To investigate the possible expression of these mitochondrial

Figure 1 Dot plot representing the alignment of a large nuclear mitochondrial DNA sequence (S12.297) with the ovine mitogenome. Sequences of the mitochondrial genome and of the nuclear mitochondrial DNA sequence are plotted on *X* axis and *Y* axis respectively. The positions indicated in the axes of the dot plot start at 1 and go to the complete length of the sequence. Therefore, dot plot representation is not on the same scale for the *X* and *Y* axes.

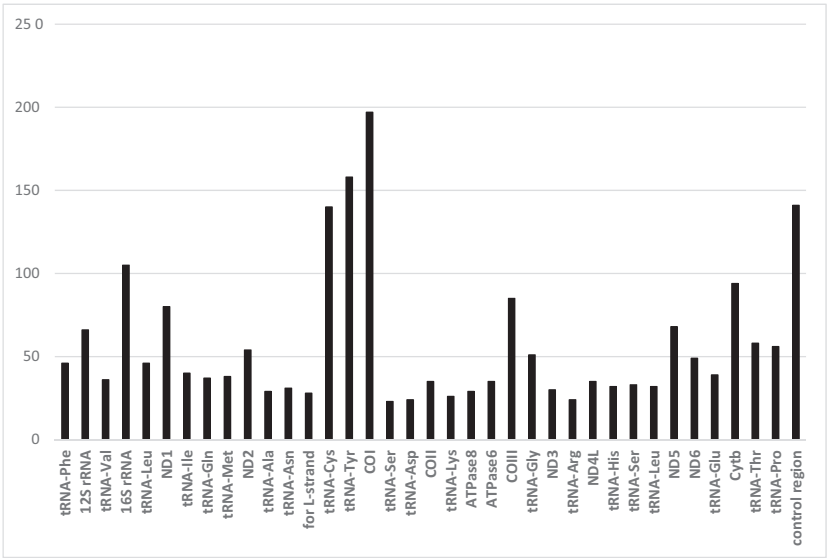

Figure 2 Number of nuclear mitochondrial DNA sequences containing, completely or partially, each region of the sheep mitogenome.

nuclear sequences, BLASTN alignments with refseq_rna and EST databases from GenBank of *Ovis aries* were performed. We excluded for these analyses NUMTs found on unplaced scaffolds. We found significant BLAST hits (*e* value threshold of 10^{-4} and identity of at least 98%) with refseq_rna sequences for three NUMT sequences. These NUMT sequences match predicted non-coding RNA sequences (*ENSOARG00020006902*, *ENSOARG00020025201* and *ENSOARG00020002007*). We also identified significant matches between the sequences of NUMTs and ESTs. However, after removing ESTs aligning perfectly to the mitochondrial genome sequence, only two NUMTs remained (Accession number EE787308.1 and DY518932.1). S3.118 and S5.171 NUMT regions show BLAST hits with ESTs. These findings suggest the possibility that these two NUMT regions might be expressed, but further investigation is needed, as only one EST per NUMT was detected and contamination of the original cDNA libraries with genomic (nuclear) DNA cannot be ruled out.

To further investigate the possible transcription of the identified NUMTs, we used the program ORF Finder (Rombel *et al.* 2002) to predict open reading frames (ORFs) located within all NUMT sequences. A total of 9201 ORFs ranging from 10 to 592 amino acids, were predicted. Among these ORFs we found 347 partial ORFs, containing a start codon but without a stop codon. Investigations showed that this is happening for 96% of the partial ORFs when the ORF is at the end of the NUMT sequences, suggesting that the remaining part of these putative ORFs might be found in the nuclear sequences flanking these NUMT regions. We further annotated the 8854 predicted full-length ORFs using the CD search tool (Marchler-Bauer & Bryant 2004) and identified 65 different conserved protein domains, within 671 ORFs from 99 NUMT regions. As expected, most of these protein domains are shared with proteins encoded by the mitochondrial genome (e.g. COX2 domain). Interestingly, we identified six different predicted ORFs, each in a different NUMT, sharing the humanin domain (Fig. S3). Humanin, first described in human, is a neuroprotective and anti-apoptotic micropeptide encoded by the 16S rRNA region of the mitochondrial genome (Nishimoto *et al.* 2004) and is not yet described in sheep. Our findings suggest the existence of nuclear-encoded *humanin* isoforms, as previously found in human (Bodzioch *et al.* 2009).

To rule out the possibility that the identified NUMTs could be artefacts of the latest ovine genome assembly, we randomly selected nine NUMT regions located into nine different chromosomes for experimental validation. PCR primers were designed within the flanking regions of the NUMT regions using PRIMER-BLAST (Ye *et al.* 2012). In addition, each primer sequence was aligned onto the mitochondrial genome sequence in order to verify its (nuclear) specificity. Primers were purchased from Integrated DNA Technologies. Primer sequences can be found in Table S3. PCRs were performed in 10 µl, using 50 ng of genomic DNA from a female sheep, as previously described (Tramontin Grau *et al.* 2020). PCR products were analysed by electrophoresis on an 0.8% agarose gel. All nine amplicons were of expected length indicating no genome assembly artefacts (Fig. S4). The nucleotide sequence of each amplicon was subsequently determined using Sanger sequencing (Eurofins Genomics). BLASTN alignment of the sequences of the amplicons to the sheep reference genome confirmed the presence of these NUMT regions.

Our study provides the first comprehensive description of the location of NUMTs in the ovine genome. We found several NUMT regions showing high similarity to the

mitochondrial genome (Table S1) that potentially could pose a risk to mitochondrial studies. For example, we discovered that NUMT region S6.180 located on chromosome 6 shares >99% identify with a mitochondrial sequence containing the *D*-loop control region, which is often used in genetic diversity and phylogenetic studies (Dymova *et al.* 2017). Special attention should therefore be given for the selection of primers for PCR amplification of mitochondrial DNA.

Acknowledgements

We are grateful to Cécile Grohs (INRAE, Jouy-en-Josas) for providing the female sheep DNA sample. We would also like to thank the two anonymous reviewers for their helpful comments. The work was supported by the National Research Institute for Agriculture, Food and Environment (INRAE).

Conflict of interest

The authors declare they have no conflict of interest.

References

Altschul S.F., Gish W., Miller W., Myers E.W. & Lipman D.J. (1990) Basic local alignment search tool. *Journal of Molecular Biology* **215**, 403–10.

Bodzioch M., Lapicka-Bodzioch K., Zapala B., Kamysz W., Kiec-Wilk B. & Dembinska-Kiec A. (2009) Evidence for potential functionality of nuclearly-encoded *humanin* isoforms. *Genomics* **94**, 247–56.

Calabrese F.M., Balacco D.L., Preste R., Diroma M.A., Forino R., Ventura M. & Attimonelli M. (2017) NumtS colonization in mammalian genomes. *Scientific Reports* **7**, 16357.

Doynova M.D., Berretta A., Jones M.B., Jasoni C.L., Vickers M.H. & O'Sullivan J.M. (2016) Interactions between mitochondrial and nuclear DNA in mammalian cells are non-random. *Mitochondrion* **30**, 87–96.

Dymova M.A., Zadorozhny A.V., Mishukova O.V., Khrapoy E.A., Druzhkova A.S., Trifonov V.A., Kichigin I.G., Tishkin A.A., Grushin S.P. & Filipenko M.L. (2017) Mitochondrial DNA analysis of ancient sheep from Altai. *Animal Genetics* **48**, 615–8.

Hazkani-Covo E., Zeller R.M. & Martin W. (2010) Molecular poltergeists: mitochondrial DNA copies (NUMTs) in sequenced nuclear genomes. *PLoS Genetics* **6**, e1000834.

Marchler-Bauer A. & Bryant S.H. (2004) CD-Search: protein domain annotations on the fly. *Nucleic Acids Research* **32**, 327–31.

Ning F.Y., Fu J. & Du Z.H. (2017) Mitochondrial DNA insertions in the nuclear *Capra hircus* genome. *Genetics and Molecular Research* **16**(1), gmr16018266. https://doi.org/10.4238/gmr16018266

Nishimoto I., Matsuoka M. & Niikura T. (2004) Unravelling the role of humanin. *Trends in Molecular Medicine* **10**, 102–5.

Rombel I.T., Sykes K.F., Rayner S. & Johnston S.A. (2002) ORF-FINDER: a vector for high-throughput gene identification. *Gene* **282**, 33–41.

Rothfuss O., Gasser T. & Patenge N. (2010) Analysis of differential DNA damage in the mitochondrial genome employing a semi-long run real-time PCR approach. *Nucleic Acids Research* **38**, e24.

Tramontin Grau E., Charles M., Féménia M., Rebours E., Vaiman A. & Rocha D. (2020) Survey of mitochondrial sequences integrated into the bovine nuclear genome. *Scientific Reports* **10**, 2077.

Yao Y.-G., Kong Q.-P., Salas A. & Bandelt H.-J. (2008) Pseudomitochondrial genome haunts disease studies. *Journal of Medical Genetics* **45**, 769–72. https://doi.org/10.1136/jmg.2008.059782

Ye J., Coulouris G., Zaretskaya I., Cutcutache I., Rozen S. & Madden T.L. (2012) Primer-BLAST: a tool to design target-specific primers for polymerase chain reaction. *BMC Bioinformatics* **13**, 134.

Supporting information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Figure S1 Distribution of NUMTs (alignment matches) and NUMT regions across all chromosomes.

Figure S2 Dot plots of highly modified NUMT regions (upper, NUMT S3.92 and bottom S25.506).

Figure S3 Alignment of sheep NUMT-derived and human mitochondrial *humanin* proteins.

Figure S4 PCR amplification of 9 NUMT regions with genomic DNA.

Table S1 List of all NUMT regions detected in the ovine genome assembly.

Table S2 NUMT regions located within gene boundaries.

Table S3 NUMT regions selected for validation.