AUTHOR VERSION

Complete mitogenomes from Kurdistani sheep: abundant centromeric nuclear copies representing diverse ancestors

Sarbast Ihsan Mustafa [a,b], Trude Schwarzacher[a] and JS Heslop-Harrison[a]

[a]*Department of Genetics and Genome Biology, University of Leicester, Leicester, U.K.*

[b] *University of Duhok, Duhok, Kurdistan Region, Iraq*

Corresponding author contact: e-mail phh4@le.ac.uk; address: Department of Genetics and Genome Biology, University of Leicester, University Road, Leicester LE1 7RH, U.K.

Short title: Mitogenomes of Kurdistani sheep

Citation: Mustafa SI, Schwarzacher T, and Heslop-Harrison JS. 2018 Complete mitogenomes from Kurdistani sheep: abundant centromeric nuclear copies representing diverse ancestors. **MITOCHONDRIAL DNA PART A, 2018 https://doi.org/10.1080/24701394.2018.1431226**

Abstract

The geographical center of domestication and species diversity for sheep (*Ovis aries*) lies around the Kurdistan region of Northern Iraq, within the 'Fertile Crescent'. From whole genome sequence reads, we assembled the mitochondrial genomes (mtDNA or mitogenome) of five animals of the two main Kurdistani sheep breeds Hamdani and Karadi and found they fitted into known sheep haplogroups (or matrilineages), with some SNPs. Haplotyping 31 animals showed presence of the main Asian (HPGA) and European (HPGB) haplogroups, as well as the rarer Anatolian haplogroup HPGC. From the sequence reads, near-complete genomes of mitochondria from wild sheep species (or subspecies), and even many sequences similar to goat (*Capra*) mitochondria, could be extracted. Analysis suggested that these polymorphic reads were nuclear mitochondrial DNA segments (*numts*). *In situ* hybridization with seven regions of mitochondria chosen from across the whole genome showed strong hybridization to the centromeric regions of all autosomal sheep chromosomes, but not the Y. Centromeres of the three submetacentric pairs and the X chromosomes showed fewer copies of *numts*, with varying abundance of different mitochondrial regions. Some mitochondrial-nuclear transfer presumably occurred before species divergence within the genus, and there has been further introgression of sheep mitochondrial sequences more recently. This high abundance of nuclear mitochondrial sequences is not reflected in the whole nuclear genome assemblies, and the accumulation near major satellite sequences at centromeres was unexpected. Mitochondrial variants including SNPs, *numts* and heteroplasmy must be rigorously validated to interpret correctly mitochondrial phylogenies and SNPs.

Keywords: nuclear mitochondrial DNA segments (*numts*);; mitogenome diversity; massively parallel sequencing; fluorescent *in situ* hybridization; sheep domestication

Introduction

Introduction Sheep were among the first group of livestock species to be domesticated, with archaeological and genetic studies showing they were farmed 8000-11000 years BP in the Fertile Crescent, covering parts of Western Asia and North-East Iraq (Ryder 1983; Zeder 2008). The Kurdistan Region in the north of Iraq corresponds to the zone of initial domestication of sheep (Figure 1 (A)). Native sheep breeds such as Hamdani and Karadi, together represent about 20% of the Iraqi sheep population (totalling 6 to 8 million head) and are restricted to the Kurdistan region; another breed, Awassi, represents 58% of all sheep in Iraq, divided between the Kurdistan and the central region (Alkass and Juma 2005; Al-Barzinj and Ali 2013). The breeds are distinctive and morphologically-well-defined, with fat-tails (Figure 1 (B) and (C); Supplementary Figure S1), and are mainly used for carpet-wool production, as well as meat and milk. Sheep breeding is one of the important sources of income for smallholder farmers in the region, and the landraces are well adapted to poor grazing habitats, showing hardiness in diverse harsh environments. Domestic sheep (*Ovis aries*) are in the Caprini tribe, Caprinae subfamily, of family Bovidae in the order Artiodactyla (even-toed ungulates). The phylogeography and classification of the wild *Ovis* species is extensively discussed (Rezaei et al. 2010), and includes the species or subspecies *O. orientalis*, *O. vignei*, *O. musimon* and *O. ammon* that occur in or close to the Kurdistan region, where the samples used in this work were collected.

Domestic sheep are reported to derive from two subspecies (Ryder 1983), with approximately five independent domestication events giving rise to modern breeds, supported by archaeological and genetic evidence (Meadows et al. 2007; Zeder 2008). Like all domesticated animals, prerequisites for domestication include changes in a number of behavioural and physiological traits such as easy management (e.g. lack of aggressiveness and flight response), environmental tolerance (hardiness and disease resistance), and productivity (for sheep, initially meat and rapid reproduction, later including wool and milk). Given their role in energy metabolism, mitochondrial genes contribute to many adaptive and productive traits, and mitochondrial genome (mtGenome, or mitogenome) variation has been associated with phenotype, including hardiness, disease tolerance and resistance, milk production and fertility (Hiendleder et al. 2008; see also MITOMAP.org in human). The definition of these specific diagnostic mutations has also improved phylogenetic information and due to their universal application, mitochondrial DNA markers are used for identification of species in food testing and archaeology (Kitpipit et al. 2014; Okuma and Hellberg 2015; Kemp et al. 2017; Nikitin et al. 2017; West et al. 2017). The population and evolutionary biology of mitochondrial DNA (mtDNA) sequences have been extensively studied in many species (Robins et al. 2007; Tillmar et al. 2013; Wani et al. 2014; Sharma et al. 2015). Nucleotide polymorphisms within mtDNA show diversity in maternal lineages and their phylogenetic relationship in different domesticated animals can be deduced or correlated with geographical locations or to resolve origins and ancestry (e.g. Kimura et al. 2010 in donkey and Yang et al. 2017 in dogs). In sheep, many authors (Meadows et al. 2007; Zhao et al. 2011; Demirci et al. 2013; Mariotti et al. 2013) amplified fragments of the mitochondrial control region (D-loop), tRNAPhe, and 12S rRNA while others (Lancioni et al. 2013; Lv et al. 2015; Hu and Gao 2016) analysed the complete mitogenomes to identify mtDNA diversity and phylogenetic

relationships. Altogether variation in mtDNA sequences has identified multiple maternal lineages by which the haplogroups of mitogenome diversity have been classified (Hiendleder et al. 1998; Pedrosa et al. 2005; Meadows et al. 2007; Meadows et al. 2011): the five major haplogroups, HPGA, HPGB, HPGC, HPGD and HPGE, are geographically wide-ranging, some being dominant and more specific to particular regions. HPGA and HPGB are most common and have been widely observed in Asia. In Europe, HPGB is considered the main maternal lineage while both HPGC and HPGE haplogroups and the less frequent HPGD have been described in Turkey, the Caucasus and China. Mitogenomes of the well-defined wild taxa (*O. ammon* and *O. vignei*) have also been sequenced (Meadows et al. 2007; Jiang and Ramachandran 2013).

It is notable that there has been minimal sampling of sheep from the south and eastern parts of the Middle East and specifically the Kurdistan region despite its central location in the fertile crescent (Lv et al., 2015, reporting a meta-analysis of sheep mitogenomes). Although the genetic diversity based on microsatellite and genetic markers of some sheep breeds of the Kurdistan region have been studied (Mohammed 2009; Al-Barzinji et al. 2011; Al-Barzinj and Ali 2013), their maternal diversity is unknown and no reports of genetic diversity of the fat-tailed sheep breeds nor their genetic significance and distinctiveness are available (Rocha et al. 2011). Many mitochondrial diversity studies have used PCR amplification of a limited number of genome regions, but reduced costs and the availability of next generation (or massively parallel) sequencing allows complete mitochondrial genome sequences to be obtained that provide a reference to study all variation in a species, not only for identifying polymorphisms which may relate to energy metabolism, but also polymorphisms that are not selectively neutral and may lie outside regions chosen for genotyping studies using universal primers. As we show here, whole mitogenome assembly can identify haplogroups, and can be exploited to develop PCR-based markers to target polymorphisms informative at levels from species through populations or breeds, to individuals.

Mitochondrial genomes are normally considered to show strictly matrilineal or maternal inheritance. Polymorphisms may be found in sequence data because of either heteroplasmy (Gorkhali et al 2016) – the occurrence of more than one mitochondrial variant (mtDNA sequence) – or the presence of nuclear mitochondrial DNA segments, *numts* (Zhang and Hewitt, 1996). Vaughan et al. (1999) could visualize incorporation of mitochondrial sequences in the chromosomes of the nuclear genome by fluorescent *in situ* hybridization. More than 200 *numts* were identified in the nuclear genome of the honeybee (Du and Qin 2015), most less than 1kbp and with identities of 75-90% to the mtDNA fragments. In mammals, Hazkani-Covo and Graur (2006) found several hundred mostly short *numts* in both human and chimpanzee including 391 orthologous *numts* present in both genomes. In general, *numts* are present in variable size and found to be highly fragmented, rearranged and distributed among and within nuclear genomes with different degrees of homology to their mtDNA sequence fragments (Zhang and Hewitt 1996; Vaughan et al. 1999; Woischnik and Moraes 2002). The abundance and mitochondrial coverage of *numts* in most species are, however, still unknown.

Here, we aimed to generate the complete mtDNA genome of the two main Kurdistani sheep landraces or breeds, Hamdani and Karadi. Using DNA from a larger panel of individuals, we aimed to identify the maternal haplogroups of the endemic breeds and the Awassi breed, more widespread in the Middle East. As a consequence of identifying multiple mitochondrial

sequence variants in each animal, we used DNA *in situ* hybridization to identify the presence and genomic location of *numts*.

Materials and methods

Sampling and isolation of genomic DNA

Blood samples including 11 males and 20 females were collected with K2E Vacutainers (Becton Dickinson) from the jugular vein of sheep, including Hamdani, Karadi and Awassi breeds during regular veterinary health checks in the University flocks, in accordance with Iraqi regulations. Individual animals were photographed to confirm breed characteristics (Figures 1 (B), (C); Tables 1 and 2; Supplementary Figure S1). Blood was sampled from flocks representing different locations of the Iraqi Kurdistan Region (Duhok, Erbil and Sulaymaniyah Governorates; see Figure 1 (A), Table 2). Total DNA, including both nuclear and mitochondrial genomes, was extracted from blood using the Wizard Genomic DNA Purification kit (Promega, Southampton, UK).

Assembly of complete mitochondrial genomes

Five samples of genomic DNA (Table 1) were sequenced using the Illumina NextSeq500 mid-throughput 2x150bp cycle system with barcoded/multiplexed total DNA samples (in two runs at the University of Florida Interdisciplinary Center for Biotechnology Research, Gainesville, FL, USA) giving 40 to 60 million reads with 5 to 6 Gb total sequence for each DNA sample. For each sample, paired end reads were assembled against the complete mtDNA genome of Oxford Down *Ovis aries* (KF938359) as a reference using low stringency (maximum mismatches per read 10%). The five complete mitogenomes were then annotated using Geneious 8.0 (Kearse et al. 2012; http://www.geneious.com). The annotated mitochondrial DNA are available in GenBank under accession numbers MF004242 - MF004246.

Data analysis and relationships

The five complete mitochondrial genomes of Hamdani and Karadi sheep breeds were aligned with published genomes from 10 domestic, 6 wild sheep species and 2 *Ovis musimon* species samples (see Supplementary Table S1 for accessions and references) representing the five main sheep haplogroups, HPGA, HPGB, HPGC, HPGD and HPGE (Hiendleder et al. 2002; Meadows et al. 2011) and used to construct a Bayesian tree. MEGA6 (Tamura et al. 2013) was used to find the best model of mitochondrial sequence alignments using the maximum likelihood criteria. Geneious software (Kearse et al. 2012; http://www.geneious.com) was used for alignment with default parameters and optimized manually. Bayesian phylogeny inference was used for analysis with MrBayes 3.2.6 (Huelsenbeck and Ronquist 2001) within Geneious and largely default parameters based on (General Time Reversible, GTR, substitution model, invariant gamma rate variation with four gamma categories, a burn-in length of 15,000 and chain length of 20,000). Trees were built for the entire mitochondrial genomes. For all analyses, the mitochondrial genome of goat (*Capra hircus*) was used as the outgroup.

PCR-RFLPS (CAPS) to survey mitochondrial sequence variation in populations

After alignment between the consensus of the main haplogroups HPGA, HPGB, HPGC, HPGD and HPGE, polymorphic sites were marked and analysed for possible restriction enzyme recognition sites to cut the DNA in one haplogroup but not the others to enable classification. Six primer pairs were designed for PCR-RFLP (CAPS; Supplementary Table S2) to span these polymorphic restriction sites, encompassing three different parts of the mtDNA genome, the ND1 gene (2850-3341nt), Cox1 gene (5437- 6024nt) and CYTB gene (14786nt-15208nt)(Supplementary Tables S3).

Genomic DNA from five sequenced and 26 additional sheep (Table 2) was amplified using the primers in 50µl total volume reaction mixture containing ddH_2O, 10x Buffer A (Kapa Biosystems, Wilmington, MA, USA), 10mM dNTP Mix (Bioline Reagents, London, UK), 10µM primers forward and reverse primers (Sigma-Aldrich, Dorset, UK) and 5U/ µL KAPA Taq DNA Polymerase (Kapa Biosystems) with 80-100ng of DNA. PCR conditions consisted of 3 min initial denaturation at 95°C, followed by 35 cycles of denaturation (95°C, 0.5min), annealing (Tm-5°C, 0.5 min) and primer extension (72°C, 1min). The final cycle added 1 min extension at 72°C followed by hold time at 16°C. Amplified fragments (2-3µl) were digested with one of eight restriction enzymes in the appropriate buffer (Supplementary Table S3) and digested PCR products were separated by gel electrophoresis (2% w/v agarose) in 1x TAE buffer [Tris base, acetic acid, EDTA (ethylene-diamine-tetra-acetic acid); containing ethidium bromide].

Sanger sequencing of polymorphic regions

Four primer pairs were designed for Sanger sequencing (Supplementary Table S2) spanning variant related positions of the ND1 gene (nt2876-3552) and Cox1 gene with part of tRNA Ser (nt6370-6916). The amplified PCR products were spanning the same region showing heterogeneity in base calls at the same positions of polymorphisms.

Variant frequencies of mitochondrial genomes

After assembly, it was evident that some sites had a substantial proportion of reads with alternative bases to the consensus. Geneious 8.0 (Kearse et al. 2012; http://www.geneious.com) was used to call variants by re-assembling all the raw reads from each of the NGS data samples to their respective annotated mitogenomes. Variant frequencies (disagreements) were configured by setting appropriate features: genetic code was set to mitochondrial vertebrates, minimum variant frequency to 0.005%, and other parameters were kept as default.

The variant frequencies were investigated separately in coding sequences (CDS) and non-coding sequences (non-CDS): total SNPs, synonymous, non-synonymous, transition and transversions were calculated (Supplementary Table S4). SNPs with extreme strand bias (top 10%) and low variant frequencies (<2.5%) were omitted from the analysis as in Guo et al. (2012).

Fluorescent in situ hybridization (FISH)

Peripheral sheep blood was collected from freshly slaughtered commercial sheep (Joseph Morris Butchers Ltd, Lutterworth, Leicestershire, UK) in sterile 50 ml tubes containing heparin. Lymphocyte short term medium contained 43.5ml of RPMI medium 1640 (1X) (Gibco™, Fisher Scientific, Loughborough, UK), 6ml of foetal calf serum and 0.5ml of antibiotic antimycotic solution (HyClone™, GE Healthcare Life Sciences, Amersham, UK, containing 10,000 U/ml penicillinG, 10,000µg/ml streptomycin, and 25µg/ml Amphotericin B). 0.5 or 0.75 ml of blood were added to 7ml medium containing 10-30µg/ml phytohemagglutinin (PHA; Sigma-Aldrich) in 5% CO2 incubator at 37ºC for 3-5 days. Metaphases were arrested by adding 50-90µl of Demecolcine solutions (10µg/ml; Sigma-Aldrich) and for further 1.5-2 hours at 37ºC. Metaphase chromosome preparations were then made as described by Schwarzacher and Heslop-Harrison (2000) using hypotonic treatment with 0.075M KCl and fixation in absolute methanol:glacial acetic acid 3:1.

Seven regions spanning the whole mitogenome were amplified from genomic DNA to use as probes for FISH (for primers see Supplementary Table S2; for conditions see above); amplification products were purified and sent to sequencing to check that they correspond to mitochondrial variants as expected. Probes except control region probe were labelled with biotin–16-dUTP (Roche Diagnostics, Burgess Hill, UK) that was labelled with digoxigenin–11-dUTP (Roche Diagnostics) using the BioPrime Array CGH random priming kit (Invitrogen™, Fisher Scientific, Loughborough, UK).

FISH followed the protocols of (Schwarzacher and Heslop-Harrison 2000). The hybridization mixture contained 50% (v/v) formamide, 20% (w/v) dextran sulphate, 2x SSC (saline sodium citrate: 0.3 M NaCl, 0.03 M sodium citrate), 50-100 ng probe, 20µg sheared salmon sperm DNA (Sigma-Aldrich), 0.3% (w/v) SDS (sodium dodecyl sulfate) and 0.12 mM EDTA. After overnight hybridization at 37°C, washes were carried out with high stringency (20% formamide and 0.1xSSC) enabling probe-target hybrids with more than 85% homology to remain stably hybridised. Biotin-labeled probes were detected with 2.0 µg/ml streptavidin conjugated to Alexa594 (Molecular Probes™, Fisher Scientific, Loughborough, UK), and digoxigenin probes were detected with 4 µg/ml anti-digoxigenin conjugated to FITC (fluorescein isothiocyanate, Roche Diagnostics). Slides were simultaneously counterstained and mounted by applying antifade mixture [6µl DAPI (4´,6-diamidino-2-phenylindole diluted in McIlvaines buffer pH 7.0; 100µg/ml) 97µl Citifluor antifade mountant solution (Citifluor™, Agar Scientific, Stansted, UK) and 97µl ddH2O]. Preparations were analyzed on a Nikon Eclipse N80i fluorescent microscope equipped with a DS-QiMc monochromatic camera (Nikon, Tokyo, Japan). Each metaphase was captured in two different filter sets: blue excitation for FITC or green excitation for Alexa594 and UV excitation for DAPI. Images were falsely colored (red for the probe and cyan for DAPI), overlaid and the contrast adjusted with NIS-Elements BR3.1 software (Nikon) using only cropping, and functions affecting the whole image equally.

Results

The mitochondrial genome in Kurdistani sheep and relationships

Total genomic DNA samples of two Hamdani, one Karadi, and one of each breed with some intermediate, partially mixed, phenotype (Table 1; Figure 1 (B) and (C); Supplementary Figure S1) were sequenced. From each of the 5 to 6 Gb of paired end reads, the complete consensus mitochondrial genomes, HamJ1, HamJ2, HamM, KarM and KarJ (Figure 2; Supplementary Figures S2-S5) were extracted by mapping to a reference *Ovis aries* mtDNA genome (KF938359 from Oxford Down; Lv et al. (2015)) and they were 16,617bp, 16,618bp or 16,619bp. The sequencing coverage was 120 to 308 times (Table 1) and is equivalent to 56 to 105 mitochondrial genomes per nuclear genome; the coverage was on average 1.46 times greater in the female than male samples. Niu et al. (2017) estimated a similar range, but lower copy numbers of mtDNA in Chinese sheep breeds using quantitative real-time PCR (qPCR). Alterations in the copy number of mtDNA have been found during cell growth and differentiation (Shay et al. 1990) and can be influenced by physiological and environmental conditions (Lee and Wei 2005).

A complement of 37 genes was found (Figure 2 and Supplementary Figures S2-S5), consisting of 22 tRNA genes, two rRNA genes (12S rRNA and 16S rRNA), 13 protein-coding genes (CDS), and one control region (D-loop), and the GC content averaged 38.9%. Variants between the five mitogenomes of the Kurdistani sheep breeds were tabulated in accordance with the sequence variant descriptions recommended by the HGVS nomenclature (Dunnen et al. 2016) generated by Mutalyzer (Supplementary Table S5). An additional tandem repeat which is found in the wild species was not present in the Kurdistani mitochondrial genomes.
[Figure 2 near here]

The phylogenetic position of the five assembled Kurdistani mitochondrial genomes was established by Bayesian tree analysis including as reference genomes published haplogroups of domestic sheep, wild sheep *O. musimon*, *O. vignei*, *O. ammon*, *O. canadensis*, and as an outgroup, goat, *Capra hircus* (Hassanin et al. 2010) (Figure 3; Supplementary Table S1). The consensus mitochondrial sequences from Kurdistan were placed on branches with the recognized sheep haplogroup HPGA (three animals of Hamdani breed) and HPGB (two animals of Karadi breed), while the other three known haplogroups HPGC, HPGD and HPGE (Meadows et al. 2011) were on separate branches. Two of the reference samples of *O. musimon* (Mouflon - HM236184, HM236185) were sisters to the haplogroup HPGB within the same subclade as the Karadi mitochondria.

PCR-RFLPs (or CAPs, Cleavage Amplification Polymorphisms) (Supplementary Table S3; Figure 4 (C) and Supplementary Figures. S6-S14) identified the major mitochondrial haplogroup in 26 male and female Hamdani, Karadi and Awassi sheep making it a total of 31 sheep analyzed in this study (Table 2 (A)). About half were HPGA with the remaining divided equally between HPGB and HPGC (Table 2(B)). As further confirmation of the presence of the three haplogroups, a primer pair was designed to amplify a part of the ND1 gene (Supplementary Table S2 (B)) and PCR products were sequenced and confirmed HPGA (H-390-P, H1a, 1Aw), HPGB (K279-P, H-368-P) and HPGC (3K 00454, K5-SUL, H-364-P).

Nuclear-mitochondrial DNA sequences (numts)

Presence of variant mitochondrial sequences

The most frequent base in the overlapping reads had been used as the consensus for each mitochondrial genome (Figure 2, and Supplementary Figures S2-S5, and described above). However, multiple sites with variant base calls were found in the sequencing reads with each animal having two or more bases (>1%) present at some positions of the sequence reads mapped to the consensus assembly. The types and frequencies of SNPs was analyzed for the mitogenome HamJ1 and selected variant regions were amplified by PCR followed by Sanger sequencing; mitogenome sequences were searched for in whole nuclear genome assemblies and used as probes for fluorescent *in situ* hybridization to mitotic chromosomes.

Characterization of variant mitochondrial regions

Within the assembly of HamJ1, a total of 394 SNPs was present in multiple, but relatively small proportion of reads (between 2.5% and 7.5% of the reads), including 262 synonymous and 23 non-synonymous SNPs within coding sequences (Supplementary Table S4). Most (363/394) were transitions while transversions were rare. Polymorphisms were present but more limited inside the control region, with 10 SNPs. Some 31 SNPs were found in each of the 12S and 16S rRNA, and 37 SNPs were found in all tRNA regions.

Figure 4(A) shows some of the 436 raw read sequences mapping to part of the ND1 gene, where 17 include variants. Another example is shown in Supplementary Figure S15 (A). PCR primers were designed to span selected polymorphic regions (Supplementary Table S2) and fragments amplified from the animals used for DNA sequencing as well as from the other sampled individuals. The PCR products were sequenced by Sanger sequencing (Figure 4(B); Supplementary Figure S15(B)); polymorphic base calls (stacked peaks from two bases) were reported only at the same locations as in the Illumina sequence reads (compare with Figure 4(A); Supplementary Figure S15(A)) indicating that next generation sequencing and assembly errors can be excluded.

Nature of variant mitochondrial regions

To assign variant reads to mitochondrial or nuclear genomes, raw reads of regions identified to contain SNPs were assembled against appropriate regions or whole consensus mitochondrial genome sequences at low stringency, allowing up to c. 10% mismatches. All the reads were then extracted and re-assembled at high stringency against the consensus mitochondrial genome for that animal. Unassembled polymorphic reads (different by more than 1% from the consensus) were extracted and used to make *de novo* assemblies (in Geneious) of these variant 150bp reads. They were then compared with the GenBank database, and the highest similarity was found to mitogenomes of several species including *O. canadensis*, *O. ammon*, *O. vignei* and genus *Capra*, and some sequences reported as nuclear including regions of *O. canadiensis* chromosome 26 and *O. aries* chromosome X. The detailed results of these comparisons, including coding and control regions and the complete assembly are shown in Supplementary Table S6. After finding the similarity with *O. ammon* and *O. vignei*, the raw reads from the

Kurdistani sheep were assembled to the mitochondrial genomes of the wild *Ovis* species; 1000 to 2000 reads showed 100% similarity.

The nuclear chromosome assemblies of *O. aries* Oar_v4.0 database showed 211 sequence fragments similar to mitochondrial sequences (total length 236,434 bp), potential nuclear mitochondrial sequences (*numts*). When the fragments were aligned back to the complete mitochondrial sequence, coverage was relatively equal over all regions with less over the control region (Figure 5); only the X chromosome (12,681bp) and chromosome 3 (8,273bp and 7,355bp) had two long assemblies covering nearly all parts of the mitochondrial genome, and 85% were less than 2kb. The 236kb of *numts* in the chromosome assemblies is 7-fold less than the total length of sequence (1,643 kb or 0.055% of all raw reads) with mitochondrial homology but different from *O. aries* mitochondria found in the raw reads.

Chromosomal localization of numts

We used fluorescent *in situ* hybridization to male metaphase sheep chromosomes with probes from seven regions of the mitochondrial genome, including coding, rRNA, and control regions, to first prove the presence of *numts*, second to localize the *numts* on the chromosomes, and third to indicate their abundance (Figure 6; Supplementary Figure S16). All seven probes gave strong *in situ* hybridization signal at the centromeres of the 23 autosomal chromosome pairs with equal to variable strength while additional signal at intercalary and subtelomeric positions could be seen occasionally (Supplementary Figure S16 (B)). This indicates that the sequences homologous to essentially all of the mitochondrial genome are integrated and amplified, potentially with degeneration as FISH conditions allow up to 15% mismatch.
[Figure 6 near here]

Discussion

Mitochondrial sequences of Kurdistani sheep

Complete sequences and abundance

Results of our study fill a gap in geographical coverage (Lv et al. 2015), with previous sheep samples from Turkey and Israel (proximate to Kurdistan) identifying multiple maternal haplogroups including HPGA, B, C, D and E (Meadows et al. 2007; Demirci et al. 2013; Rafia and Tarang 2016). Our study indicates that sheep breeds from the Kurdistan Region of Iraq originate from multiple maternal lineages within known consensus diversity, notwithstanding the geographic location of Kurdistan near the center of sheep diversity and domestication.

Complete mitochondrial genome sequences were assembled from five sheep of two most widespread breeds, Hamdani and Karadi (Figure 1 and Supplementary Figure S1). These fitted to the known sheep HPGA and HPGB haplogroups (Table 1 and Figure 3), but had additional previously undescribed SNPs (Supplementary Tables S1 and S4).
The blood lymphocyte DNA used here was suitable for extracting and assembling whole mitochondrial sequences. As reported by Ding et al. (2015) in human, the female samples had more copies of mitochondria per nuclear genome cell than male cells (Table 1). While there is only limited energy metabolism in blood lymphocytes, once growth has finished, females,

because of pregnancy and lactation, may have a higher energy requirement and hence additional mitochondria, and female mammals usually have longer lifespan.

Kurdistani breed haplogroups

The mitochondrial consensus haplogroups of the three Kurdistani sheep breeds sampled here fitted within three of the five main *Ovis aries* haplogroups known to occur in European and Asian breeds, HPGA, HPGB and HPGC; no HPGD and HPGE were found in the samples analyzed. Half of the Kurdistani sheep were HPGA, with the others HPGB and HPGC. In all breeds surveyed previously, the most common Asian haplogroup has been reported as HPGA, while most European sheep have HPGB (Wood and Phua 1996; Hiendleder et al. 1998), consistent with our findings in Kurdistan (west Asia). Lineage HPGC, found here, has been observed in fat-tail Asian and middle Eastern sheep breeds (Pedrosa et al. 2005) and is frequent in sheep from Southeast Anatolia (Demirci et al. 2013). Rafia and Tarang (2016) looked at Iranian breeds suggesting gene flow and intermixing while Niu et al. (2017) found the three main maternal lineages HPGA, HPGB and HPGC further East in Tibetan sheep.

With respect to separation of lineages, HPGB represents the mouflon / *Ovis musimon* domestication event while HPGA originates from the *O. aries* lineage. Hiendleder et al. (1998) estimates that these European- and Asian-types of haplogroup separated 375,000 to 750,000 years ago. Based on the polymorphisms seen here, the separation between our haplogroup sequence variants HPGA and HPGB is estimated as occurring similarly 400,000 to 800,000 years ago. Loftus et al. (1994) showed a similar time of separation of zebu and taurine cattle (0.2 to 1 million years ago; 8% different in hypervariable mitochondrial region), both events being well before domestication (c. 10,000 years ago). No additional haplogroups, not described for domestic sheep before were found, despite geographic proximity to wild sheep species. This is different of the situation in cattle where of eleven bison herds in the US, one has cattle mitochondria (Halbert and Derr 2006), an intergeneric transfer of mitogroups. The sheep here, from the interface of Europe and Asia, are unsurprisingly mixtures of the haplogroups.

Nuclear mitochondrial sequences, numts

Examination of the raw read assembly showed that bases were different to the consensus mitochondrial sequence (Figure 4; Supplementary Figure S15). Many such differences are ascribed to sequencing errors which are removed by the high coverage. However, sites were noted where multiple raw reads, from both forward and reverse directions, showed the same alternative bases. This could arise from artefacts including 1) systematic instrument/chemistry/base calling errors; 2) wrong assembly of duplications in the mitochondrial sequence; 3) mapping of nuclear copies of homologous sequences to the mitochondrial genome; or 4) systematic of chance accumulations of sequencing errors. Continuous improvement of base calling algorithms, chemistry or protocols, read-lengths, and instrument software mean that the rates and nature of errors in Illumina sequence calls are continuously changing and not systematically documented, so reference to error rates in published works such as Minoche et al. (2011) or Wall et al. (2014) is of little value. The sequencing here was carried out in two multiplex NextSeq500 mid-throughput 2x150bp runs, with two sheep and three samples of plants, and three sheep and three plant samples,

respectively. The plant reads gave c. 40-fold coverage of the cytoplasmic chloroplast genome of *Taraxacum* (Salih et al. 2017) which was assembled with a similar approach to that described here. The overall error rate in the sequence calls (<1%) was identical to that in the sheep data omitting the alternative calls, and in particular, no systematic alternative reads were seen in other high-coverage assemblies, suggesting that the multiple reads of different sequences here are not sequencing artefacts.

Whole genome assembly algorithms are not optimized to assemble *numts* in chromosomes. BAC sequencing (now largely historical) would include the whole 16kb mt genomes within their typical 100kb (although multiple tandemly arranged copies, and recombination or chimerism would be hard to rule out). Our whole genome read analysis shows the abundance of non-*O.aries* mitochondria (0.05% of the whole nuclear genome). It will not detect which reads of the *O. aries*-type come from nuclear copies *vs* mitochondria.

As shown in the BLAST searches (Supplementary Table S6), the whole genome assembly does not include even non-*O. aries* type *numts*: only a few mitochondrial sequences are assembled on 3 or 4 chromosomes, far from the signal from multiple copies of the mt genome on every one of the acrocentric autosomes after FISH to metaphase chromosome preparations (see Figure 6 and Supplementary Figure S16), or from the high level of polymorphisms seen in the PCR results. In the analysis, here, 0.055% of all reads, representing 1.6Mb of DNA per genome (or 100 mitochondrial genome copies), were homologous to mitochondrial sequences other than the consensus type. The strong signal on all the 23 pairs of autosomes chromosomes, from all domains of the mitochondrial genome, suggests more than an average of four copies of the mitochondrial genome per chromosome, so it is likely that some of the consensus reads also originate from nuclear copies. Hazkani-Covo and Graur (2006) have suggested there are 300 to 400 *numts* in human and chimpanzee; cats also have been reported to have *numts*, and Vaughan et al. (1999) have detected abundant mitochondrial sequences on chromosomes in the grasshopper *Chorthippus*. Amplification of large mtDNA fragments longer than 2kbp as *numts* have been found in many metazoan in short fragments (less than 1000bp) (Bensasson et al. 2001).

Fluorescent *in situ* hybridization results show very little *numt* signal on sub-metacentric chromosome pairs, with only some weak signal on chromosome 1. Submetacentrics and acrocentrics differ in satellite organization (Chaves et al. 2003) and it will be interesting to determine how satellite sequences are organized with respect to *numts* in centromeric regions of acrocentrics and the evolutionarily fused autosomal arms in sheep submetacentric chromosomes, and also in Bovidae species with different numbers of submetacentric chromosomes. Miraldo et al. (2012) shows how *numts* were transferred before the separation of the extant species of lizard. In our data, the sheep include mitochondrial sequences similar to other species of *Ovis* and even *Capra* (Supplementary Table S6), suggesting ancient transfer before separation of the species of genera.

Evolutionarily, there is a trend for organellar genes to be transferred to the nucleus where the gene is functional in encoding proteins, and normally the gene acquires features of the nuclear genome such as the nuclear gene code (including via exon shuffling), promoters and transit peptides (Wischmann and Schuster 1995; Gunbin et al. 2017). The *numts* recognized here retained the sequences of other species and all polymorphisms corresponded

to sites previously reported to vary between *Ovis* sequences; it is unknown whether they have any transcriptional activity.

There are sporadic reports of mitochondrial heteroplasmy in vertebrates arising from mutations (e.g. in human; Wallace and Chalkia 2013; Stewart and Chinnery 2015) and from biparental inheritance (e.g. in great tit; Kvist et al. 2003). In sheep, Zhao et al. (2004) reported paternal as well as maternal (biparental) inheritance of mitochondria, although this has not been found in normal offspring of other mammals including human (Pyle et al. 2015). The mitochondrial DNA composition of seven foetuses and five lambs cloned from foetal fibroblasts showed heteroplasmy in seven of twelve clones tested (Burgstaller et al. 2007). Meadows et al. (2011) quoting and extending (Hiendleder et al. 2002) removed the repeat unit of the mitochondrial control region from phylogenies because of its known heteroplasmic behaviour, and Meadows et al. (2007) state 'Others have noted that low-frequency mtDNA haplogroups such as HD and HE [HPGD and HPGE] may in fact be nuclear mitochondrial pseudogenes (*numts*) (Parr et al. 2006)' but consider *numts* from these haplogroups extremely unlikely by analysing control regions. Nevertheless, Parr et al. (2006) were able to clone the full-length mitochondrial genome from nuclear DNA.

Conclusions

Analysis of the mitochondrial DNA sequences of fat-tailed Kurdistani sheep, sampled from the center of diversity and domestication, were found to have three of the five major haplogoups of the domestic species. The presence of both the Asian and European types support the multiple domestication events, and the diversity suggests ongoing gene flow, presumably occurring as sheep are traded and move, on top of the historical waves of distribution. The HPGC haplogroup found is likely to represent recent introgression from wild species rather than independent domestication as HPGC follows the geographic range of fat-tailed breeds (Lv et al. 2015). While no entirely novel haplogroups were identified in this study, nor was there evidence for ongoing introgression, some SNP variants may represent important genotypes for conservation of genetic diversity in sheep and be resources for geneticists and breeders of the strong and robust fat-tailed types.

The whole genome sequencing results showed presence of substantial numbers of copies of essentially the whole mitochondrial genomes of other species of *Ovis* and *Capra*. *In situ* hybridization showed the mitochondrial genome was present on nuclear chromosomes, particularly at the centromeres of the acrocentric autosomes. These *numts* were presumably introgressed before separation of the modern species, over 4 MY. But the strong *in situ* signals indicate presence of the *O. aries* mitochondrial sequences too, although the chromosome assemblies show very few *numts*. Our results emphasize the need for considering *numts* in analysis of phylogeny (and heteroplasmy or identification of mitochondrial disease variants), and also the need to improve genome assembly algorithms to account for repetitive sequences, including mitochondrial-related *numts*.

Supplementary Material

Supplementary data are available at Mitochondrial DNA Part A. (see separate file)

Data Availability

The annotated mitochondrial DNA sequences are available in GenBank under accession numbers MF004242, MF004243, MF004244, MF004245 and MF004246.

Geolocation information

Blood and DNA samples for the present study were collected from different sheep farms originating from different geographical locations in Iraqi Kurdistan Region. Coordinates for Duhok Governorate are (36.8679N, 42.9488W); Erbil Governorate (36.1911N, 44.0091W); and Sulaymaniyah Governorate (35.5641N, 45.3756W), Iraq.

Acknowledgements

We would like to thank Prof. Jaladet Mohammed Saleh and Dilan Jasim Khalil for letting us to use their laboratory facilities in the Research Center at the University of Duhok for DNA extraction. We thank Joseph Morris Butchers Ltd, Lutterworth for blood samples for short-term lymphocyte cultures. We also thank Ashti I. Abdulrahman for producing the map in Figure 1.

Disclosure statement

The authors report no conflicts of interest. The authors alone are responsible for the content and writing of the paper.

Funding

This project was supported by a PhD scholarship (HCDP; Human Capacity Development Program) by the Kurdistan Regional Government-Iraq to Sarbast Mustafa.

ORCID

JS Heslop-Harrison; https://orcid.org/0000-0002-3105-2167

Trude Schwarzacher; https://orcid.org/0000-0001-8310-5489

References

Al-Barzinj YMS, Ali MK. 2013. Genetic diversity among some sheep breeds in Sulaimani Governorate using RAPD-PCR Technique. J Life Sci. 7:971-979.

Al-Barzinji YMS, Lababidi S, Rischkowsky B, Al-Rawi A, Tibbo M, Hassen H, Baum M. 2011. Assessing genetic diversity of Hamdani sheep breed in Kurdistan region of Iraq using microsatellite markers. African J Biotech. 10:15109-15116.

Alkass JE, Juma KH. 2005. Small ruminant breeds of Iraq. In: Characterization of small ruminant breeds in West Asia and North Africa. Small Ruminants Research 1:63-101.

Bensasson D, Zhang D-X, Hartl DL, Hewitt GM. 2001. Mitochondrial pseudogenes: evolution's misplaced witnesses. Trends Ecol Evol. 16:314-321.

Burgstaller JP, Schinogl P, Dinnyes A, Müller M, Steinborn R. 2007. Mitochondrial DNA heteroplasmy in ovine fetuses and sheep cloned by somatic cell nuclear transfer. BMC Dev Biol. 7:141.

Chaves R, Adega F, Wienberg J, Guedes-Pinto H, Heslop-Harrison JS. 2003. Molecular cytogenetic analysis and centromeric satellite organization of a novel 8;11 translocation in sheep: a possible intermediate in biarmed chromosome evolution. Mammalian Genome 14:706-710.

Demirci S, Baştanlar EK, Dağtaş ND, Pişkin E, Engin A, Özer F, Yüncü E, Doğan ŞA, Togan İ. 2013. Mitochondrial DNA diversity of modern, ancient and wild sheep (*Ovis gmelinii anatolica*) from Turkey: new insights on the evolutionary history of sheep. PloS One. 8:e81952.

Ding J, Sidore C, Butler TJ, Wing MK, Qian Y, Meirelles O, Busonero F, Tsoi LC, Maschio A, Angius A. 2015. Assessing mitochondrial DNA variation and copy number in lymphocytes of ~ 2,000 Sardinians using tailored sequencing analysis tools. PLoS Genetics. 11:e1005306.

Du W, Qin Y. 2015. Distribution of mitochondrial DNA fragments in the nuclear genome of the honeybee. Genet Mol Res. 14:13375-13379.

Dunnen JT, Dalgleish R, Maglott DR, Hart RK, Greenblatt MS, McGowan-Jordan J, Roux AF, Smith T, Antonarakis SE, Taschner PE. 2016. HGVS recommendations for the description of sequence variants: 2016 Update. Human Mutation. 37:564-569.

Gorkhali NA, Jiang L, Shrestha BS, He XH, Junzhao Q, Han JL, Ma YH. 2016. High occurrence of mitochondrial heteroplasmy in nepalese indigenous sheep (*Ovis aries*) compared to Chinese sheep. Mitochondrial DNA Part A. 3: 27:2645-2647

Gunbin K, Peshkin L, Popadin K, Annis S, Ackermann RR, Khrapko K. 2017. Integration of mtDNA pseudogenes into the nuclear genome coincides with speciation of the human genus. A hypothesis. Mitochondrion. 34:20-23.

Guo Y, Li J, Li C-I, Long J, Samuels DC, Shyr Y. 2012. The effect of strand bias in Illumina short-read sequencing data. BMC genomics. 13:666.

Halbert ND, Derr JN. 2006. A comprehensive evaluation of cattle introgression into US federal bison herds. J Heredity. 98:1-12.

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.

Hazkani-Covo E, Graur D. 2006. A comparative analysis of numt evolution in human and chimpanzee. Mol Biol Evol. 24:13-18.

Hiendleder S, Kaupe B, Wassmuth R, Janke A. 2002. Molecular analysis of wild and domestic sheep questions current nomenclature and provides evidence for domestication from two different subspecies. Proc Roy Soc Lond B. 269:893-904.

Hiendleder S, Lewalski H, Janke A. 2008. Complete mitochondrial genomes of *Bos taurus* and *Bos indicus* provide new insights into intra-species variation, taxonomy and domestication. Cytogenet Genome Res. 120:150-156.

Hiendleder S, Lewalski H, Wassmuth R, Janke A. 1998. The complete mitochondrial DNA sequence of the domestic sheep (*Ovis aries*) and comparison with the other major ovine haplotype. J Mol Evol. 47:441-448.

Hiendleder S, Mainz K, Plante Y, Lewalski H. 1998. Analysis of mitochondrial DNA indicates that domestic sheep are derived from two different ancestral maternal sources: no evidence for contributions from Urial and Argali sheep. J Heredity. 89:113-120.

Hu X-d, Gao L-z. 2016. The complete mitochondrial genome of domestic sheep, Ovis aries. Mitochondrial DNA Part A. 27:1425-1427.

Huelsenbeck JP, Ronquist F. 2001. MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics. 17:754-755.

Jiang S-Y, Ramachandran S. 2013. Genome-wide survey and comparative analysis of LTR retrotransposons and their captured genes in rice and sorghum. PloS One. 8:e71118.

Kearse M, Moir R, Wilson A, Stones-Havas S, Cheung M, Sturrock S, Buxton S, Cooper A, Markowitz S, Duran C. 2012. Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics. 28:1647-1649.

Kemp BM, Judd K, Monroe C, Eerkens JW, Hilldorfer L, Cordray C, Schad R, Reams E, Ortman SG, Kohler TA. 2017. Prehistoric mitochondrial DNA of domesticate animals supports a 13th century exodus from the northern US southwest. PloS one. 12:e0178882.

Kimura B, Marshall FB, Chen S, Rosenbom S, Moehlman PD, Tuross N, Sabin RC, Peters J, Barich B, Yohannes H. 2010. Ancient DNA from Nubian and Somali wild ass provides insights into donkey ancestry and domestication. Proceedings of the Royal Society B. doi. 278:50–57.

Kitpipit T, Sittichan K, Thanakiatkrai P. 2014. Direct-multiplex PCR assay for meat species identification in food products. Food Chemistry. 163:77-82.

Kvist L, Martens J, Nazarenko AA, Orell M. 2003. Paternal leakage of mitochondrial DNA in the great tit (*Parus major*). Mol Biol Evol. 20:243-247.

Lancioni H, Di Lorenzo P, Ceccobelli S, Perego UA, Miglio A, Landi V, Antognoni MT, Sarti FM, Lasagna E, Achilli A. 2013. Phylogenetic relationships of three Italian Merino-derived sheep breeds evaluated through a complete mitogenome analysis. PloS one. 8:e73712.

Lee HC, Wei YH. 2005. Mitochondrial biogenesis and mitochondrial DNA maintenance of mammalian cells under oxidative stress. Int J Biochem Cell Biol. 37:822–834.

Loftus RT, MacHugh DE, Bradley DG, Sharp PM, Cunningham P. 1994. Evidence for two independent domestications of cattle. Proc Nat Acad Sci. 91:2757-2761.

Lv F-H, Peng W-F, Yang J, Zhao Y-X, Li W-R, Liu M-J, Ma Y-H, Zhao Q-J, Yang G-L, Wang F. 2015. Mitogenomic meta-analysis identifies two phases of migration in the history of eastern Eurasian sheep. Mol Biol Evol. 32:2515-2533.

Mariotti M, Valentini A, Marsan PA, Pariset L. 2013. Mitochondrial DNA of seven Italian sheep breeds shows faint signatures of domestication and suggests recent breed formation. Mitochondrial DNA. 24:577-583.

Meadows JR, Hiendleder S, Kijas JW. 2011. Haplogroup relationships between domestic and wild sheep resolved using a mitogenome panel. Heredity. 106:700-706.

Meadows JR, Cemal I, Karaca O, Gootwine E, Kijas JW. 2007. Five ovine mitochondrial lineages identified from sheep breeds of the near East. Genetics. 175:1371-1379.

Minoche AE, Dohm JC, Himmelbauer H. 2011. Evaluation of genomic high-throughput sequencing data generated on Illumina HiSeq and genome analyzer systems. Genome Biol. 12:R112.

Miraldo A, Hewitt GM, Dear PH, Paulo OS, Emerson BC. 2012. Numts help to reconstruct the demographic history of the ocellated lizard (*Lacerta lepida*) in a secondary contact zone. Mol Ecol. 21:1005-1018.

Mohammed A. 2009. Genetic diversity in some Iraqi sheep breeds using molecular techniques. PhD Thesis, University of Duhok, Iraq.

Nikitin AG, Potekhina I, Rohland N, Mallick S, Reich D, Lillie M. 2017. Mitochondrial DNA analysis of eneolithic trypillians from Ukraine reveals neolithic farming genetic roots. PloS one. 12:e0172952.

Niu L, Chen X, Xiao P, Zhao Q, Zhou J, Hu J, Sun H, Guo J, Li L, Wang L and Zhang H. 2017. Detecting signatures of selection within the Tibetan sheep mitochondrial genome. Mitochondrial DNA Part A, 28:801-809.

Okuma TA, Hellberg RS. 2015. Identification of meat species in pet foods using a real-time polymerase chain reaction (PCR) assay. Food Control. 50:9-17.

Parr RL, Maki J, Reguly B, Dakubo GD, Aguirre A, Wittock R, Robinson K, Jakupciak JP, Thayer RE. 2006. The pseudo-mitochondrial genome influences mistakes in heteroplasmy interpretation. BMC Genomics. 7:185.

Pedrosa S, Uzun M, Arranz J-J, Gutiérrez-Gil B, San Primitivo F, Bayón Y. 2005. Evidence of three maternal lineages in Near Eastern sheep supporting multiple domestication events. Proc Roy Soc Lond B. 272:2211-2217.

Pyle A, Hudson G, Wilson IJ, Coxhead J, Smertenko T, Herbert M, Santibanez-Koref M, Chinnery PF. 2015. Extreme-depth re-sequencing of mitochondrial DNA finds no evidence of paternal transmission in humans. PLoS Genetics. 11:e1005040.

Rafia P, Tarang A. 2016. Sequence Variations of Mitochondrial DNA Displacement-Loop in Iranian Indigenous Sheep Breeds. Iranian J Appl Anim Sci. 6:363-368.

Rezaei HR, Naderi S, Chintauan-Marquier IC, Taberlet P, Virk AT, Naghash HR, Rioux D, Kaboli M, Pompanon F. 2010. Evolution and taxonomy of the wild species of the genus *Ovis* (Mammalia, Artiodactyla, Bovidae). Mol Phylogenet Evol. 54:315-326.

Robins JH, Hingston M, MATISOO-SMITH E, Ross HA. 2007. Identifying *Rattus* species using mitochondrial DNA. Molecular Ecology Resources. 7:717-729.

Rocha J, Chen S, Beja-Pereira A. 2011. Molecular evidence for fat-tailed sheep domestication. Trop Anim Health Prod. 43:1237-1243.

Ryder ML. 1983. Sheep and man. London (UK): Duckworth.

Salih RHM, Majeský Ľ, Schwarzacher T, Gornall R, Heslop-Harrison P. 2017. Complete chloroplast genomes from apomictic Taraxacum (Asteraceae): Identity and variation between three microspecies. PloS One. 12:e0168008.

Schwarzacher T, Heslop-Harrison P. 2000. Practical in situ hybridization. Oxford (UK) BIOS Scientific Publisher.

Sharma R, Kishore A, Mukesh M, Ahlawat S, Maitra A, Pandey AK, Tantia MS. 2015. Genetic diversity and relationship of Indian cattle inferred from microsatellite and mitochondrial DNA markers. BMC genetics. 16:73.

Shay JW, Pierce DJ, Werbin H. 1990. Mitochondrial DNA copy number is proportional to total cell DNA under a variety of growth conditions. J Biol Chem. 265:14802–14807.

Stewart JB, Chinnery PF. 2015. The dynamics of mitochondrial DNA heteroplasmy: implications for human health and disease. Nature Rev Genet. 16:530-542.

Tamura K, Stecher G, Peterson D, Filipski A, Kumar S. 2013. MEGA6: molecular evolutionary genetics analysis version 6.0. Mol Biol Evol. 30:2725-2729.

Tillmar AO, Dell'Amico B, Welander J, Holmlund G. 2013. A universal method for species identification of mammals utilizing next generation sequencing for the analysis of DNA mixtures. PloS one. 8:e83761.

Vaughan H, Heslop-Harrison J, Hewitt G. 1999. The localization of mitochondrial sequences to chromosomal DNA in orthopterans. Genome. 42:874-880.

Wall JD, Tang LF, Zerbe B, Kvale MN, Kwok P-Y, Schaefer C, Risch N. 2014. Estimating genotype error rates from high-coverage next-generation sequence data. Genome Res. 24:1734-1739.

Wallace DC, Chalkia D. 2013. Mitochondrial DNA genetics and the heteroplasmy conundrum in evolution and disease. Cold Spring Harb Perspect Biol. 5:a021220.

Wani CE, Yousif IA, Ibrahim ME, Musa HH. 2014. Molecular characterization of Sudanese and southern Sudanese chicken breeds using mtDNA D-loop. Genet Res Int. 2014:928420.

West C, Hofman CA, Ebbert S, Martin J, Shirazi S, Dunning S, Maldonado JE. 2017. Integrating archaeology and ancient DNA analysis to address invasive species colonization in the Gulf of Alaska. Conservation Biology. 31:1163-1172.

Wischmann C, Schuster W. 1995. Transfer of rps10 from the mitochondrion to the nucleus in Arabidopsis thaliana: evidence for RNA-mediated transfer and exon shuffling at the integration site. FEBS Letters. 374:152-156.

Woischnik M, Moraes CT. 2002. Pattern of organization of human mitochondrial pseudogenes in the nuclear genome. Genome Res. 12:885-893.

Wood N, Phua S. 1996. Variation in the control region sequence of the sheep mitochondrial genome. Anim Genet. 27:25-33.

Yang H, Wang G, Wang M, Ma Y, Yin T, Fan R, Wu H, Zhong L, Irwin DM, Zhai W. 2017. The origin of chow chows in the light of the East Asian breeds. BMC Genomics. 18:174.

Zeder MA. 2008. Domestication and early agriculture in the Mediterranean Basin: Origins, diffusion, and impact. Proc Natl Acad Sci. 105:11597-11604.

Zhang D-X, Hewitt GM. 1996. Nuclear integrations: challenges for mitochondrial DNA markers. Trends Ecol Evol. 11:247-251.

Zhao X, Li N, Guo W, Hu X, Liu Z, Gong G, Wang A, Feng J, Wu C. 2004. Further evidence for paternal inheritance of mitochondrial DNA in the sheep (*Ovis aries*). Heredity. 93:399-403.

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.

Table 1. Breed classification based on morphological traits and mitochondrial genome assembly data of the five Kurdistan sheep using Illumina NextSeq500 of total genomic DNA. Some individuals ('mixed') show crossbreed characters since there is no controlled pedigree breeding; the widespread Awassi breed is poorly defined.

Sample code	**Morphological traits[a)] /sex**	**Maternal haplogroup[b)]**	**Mitogenome and GenBank accession**	**Mitogenome size**	**Assembled reads (n)**	**Coverage [d)]**
Hb4	Hamdani / M	HPGA	HamJ1_ MF004243	16618 bp	20866	188
H115-P	Hamdani mixed with Karadi / M	HPGA	HamJ2_ MF004242	16619 bp	25715	232
H369-P	Hamdani / M	HPGA	HamM_ MF004244	16618 bp	13269	120
K279-P	Karadi / M	HPGB	KarM_ MF004246	16617 bp	34117	308
5546	Karadi mixed with Awassi / F	HPGB	KarJ_ MF004245	16617 bp	16905	153

NOTE – M, male; F, female; n, number

a- Based on size, tail, ear and wool characteristics, see Figure 1B,C and Supplementary Figure S1.

b- Haplogroups were identified based on the phylogenetic relation using known haplogroups from Meadows et al (2011); see Figure 3.

c- Coverage = Number of assembled reads*150 [average read length] /mitogenome size

Table 2. Maternal haplogroups (A) and their frequencies (B) of 31 Kurdistani sheep determined by genomic NGS data and genotyping by PCR-RFLP.

(A)

Breed[a)]	Sample location [b)]	Sample code/ Mitogenome [c)]	Sex	Haplogroup [d)]
Hamdani	Duhok	H-364-P	F	HPGC
Hamdani	**Erbil**	**H-369-P/HamM [c)]**	**M**	**HPGA**
Hamdani	Duhok	H-368-P	F	HPGB
Hamdani	Duhok	H-390-P	F	HPGA
Hamdani	Duhok	H-374-P	F	HPGB
Hamdani	Duhok	H1a	M	HPGA
Hamdani	Erbil	Hb2-a	F	HPGC
Hamdani	Erbil	Hb1-B	F	HPGA
Hamdani	**Duhok**	**H115-P/HamJ2 [c)]**	**M**	**HPGA**
Hamdani	**Erbil**	**Hb4/HamJ1 [c)]**	**M**	**HPGA**
Karadi	**Erbil**	**K279-P/KarM [c)]**	**M**	**HPGB**
Karadi	Duhok	K972-P	F	HPGA
Karadi	Duhok	K970-P	F	HPGB
Karadi	Duhok	K680-P	F	HPGC
Karadi	Duhok	K688-P	F	HPGC
Karadi	Duhok	K1a	M	HPGB
Karadi	Sulaymaniyah	K5-SUL	M	HPGC
Karadi	Sulaymaniyah	K6-SUL	M	HPGC
Karadi	Sulaymaniyah	K7-SUL	M	HPGB
Karadi	Sulaymaniyah	K8-SUL	M	HPGA
Karadi	Duhok	KB5-B	F	HPGA
Karadi	Duhok	KB3	F	HPGC
Karadi	Duhok	1K	M	HPGA
Karadi	Duhok	2K – 5350	F	HPGB
Karadi	Duhok	3K 00454	F	HPGC
Karadi	Duhok	4K-5530	F	HPGA
Karadi	**Duhok**	**5546/KarJ [c)]**	**F**	**HPGB**
Awassi	Duhok	1Aw	F	HPGA
Awassi	Duhok	2Aw	F	HPGA
Awassi	Duhok	4Aw	F	HPGB
Awassi	Duhok	5Aw	F	HPGA

(B)

Breed[a)]	Sample size	Haplogroup[d)]				
		HPGA	HPGB	HPGC	HPGD	HPGE
Hamdani	9	5	2	2	0	0
Karadi	18	6	6	6	0	0
Awassi	4	3	1	0	0	0
Total	31	14	9	8	0	0
Frequency %		**45.2**	**29.0**	**25.8**	**0**	**0**

NOTE – F, female; M, male

a) Based on predominant phenotypic characteristics, see foot note a) Table 1.

b) Coordinates for Duhok Governorate are 36.8679N, 42.9488W; Erbil Governorate 36.1911N, 44.0091W; Sulaymaniyah Governorate (35.5641N, 45.3756W)

c) Mitogenomes were assembled from 5 target sheep DNA samples; see Table1. Haplogroups as defined by Meadows et al (2011), determined by genomic NGS data [in bold, see c)] or PCR-RFLP (see Supplementary Table S4 and Supplementary Figures S6 to S14)

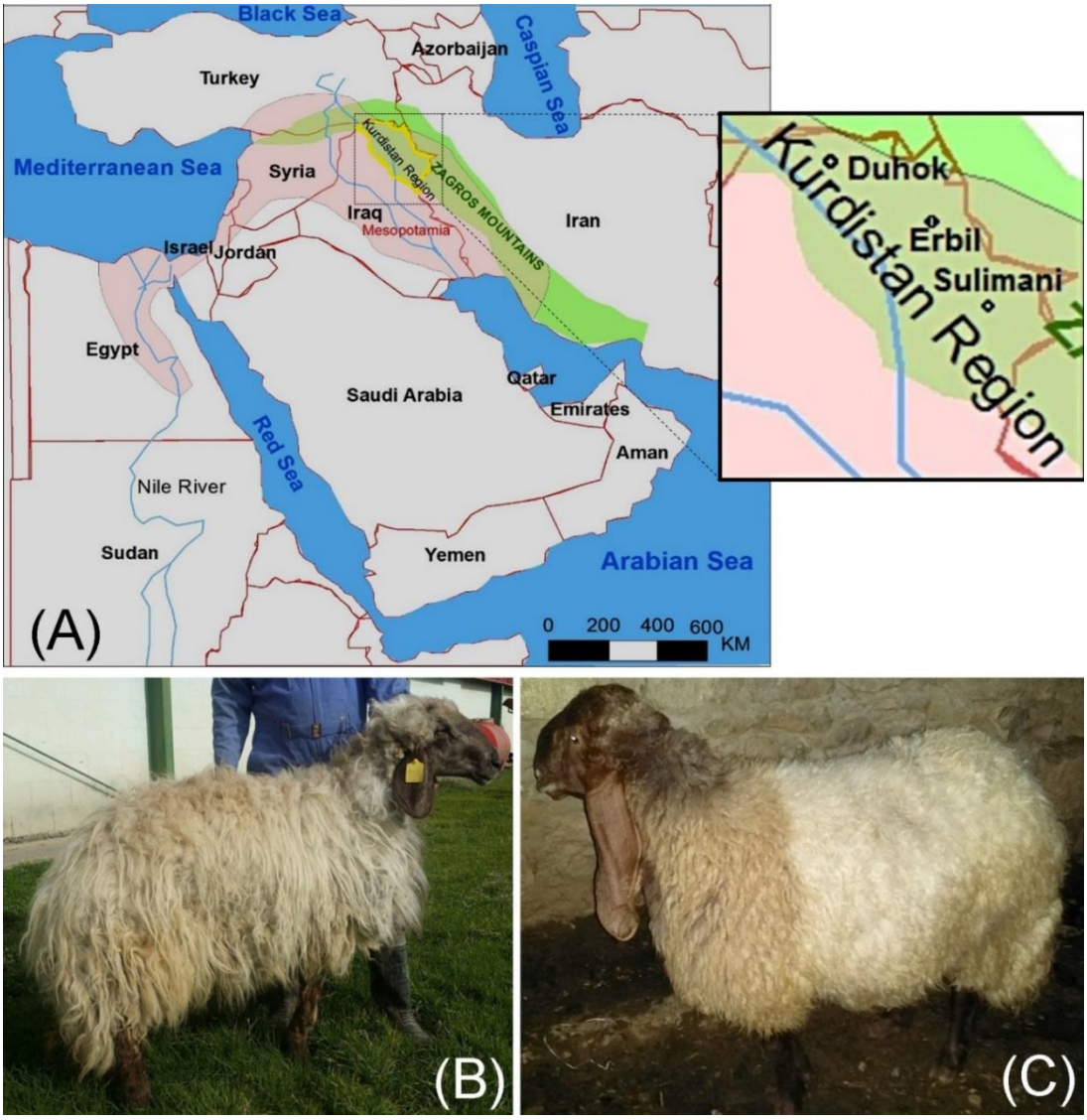

Figure 1. Locations of sampling and breed characteristics of Kurdistani sheep. (A) DNA was sampled from 31 sheep of Karadi, Hamdani and Awassi sheep in the Iraqi Kurdistan region (Duhok, Erbil and Sulaymaniyah governorates) in the 'Fertile Crescent' (pink shading) with high species and genotype diversity, and domestication events. (Map based on Arc GIS V10.3, prepared by Ashti I. Abdulrahman; for sheep collection see Supplementary Table S1). (B) and (C) Breed characteristics of Kurdistani sheep include fat tail, long wool and long ears, Roman nose and specific coloration. Karadi sheep (B), showing ram K279-P, tend to have yellowish very coarse wool, black faces and long ears while Hamdani sheep (C) showing ram Hb4, are larger and have longer ears than Karadi sheep. Tails almost reach the ground and are characterized by an outward-curving point; their fleece is more whitish but often speckled. The Awassi breed (see Supplementary Figure S1) are commonly white with red to brown faces and produce carpet quality wool; they are often horned

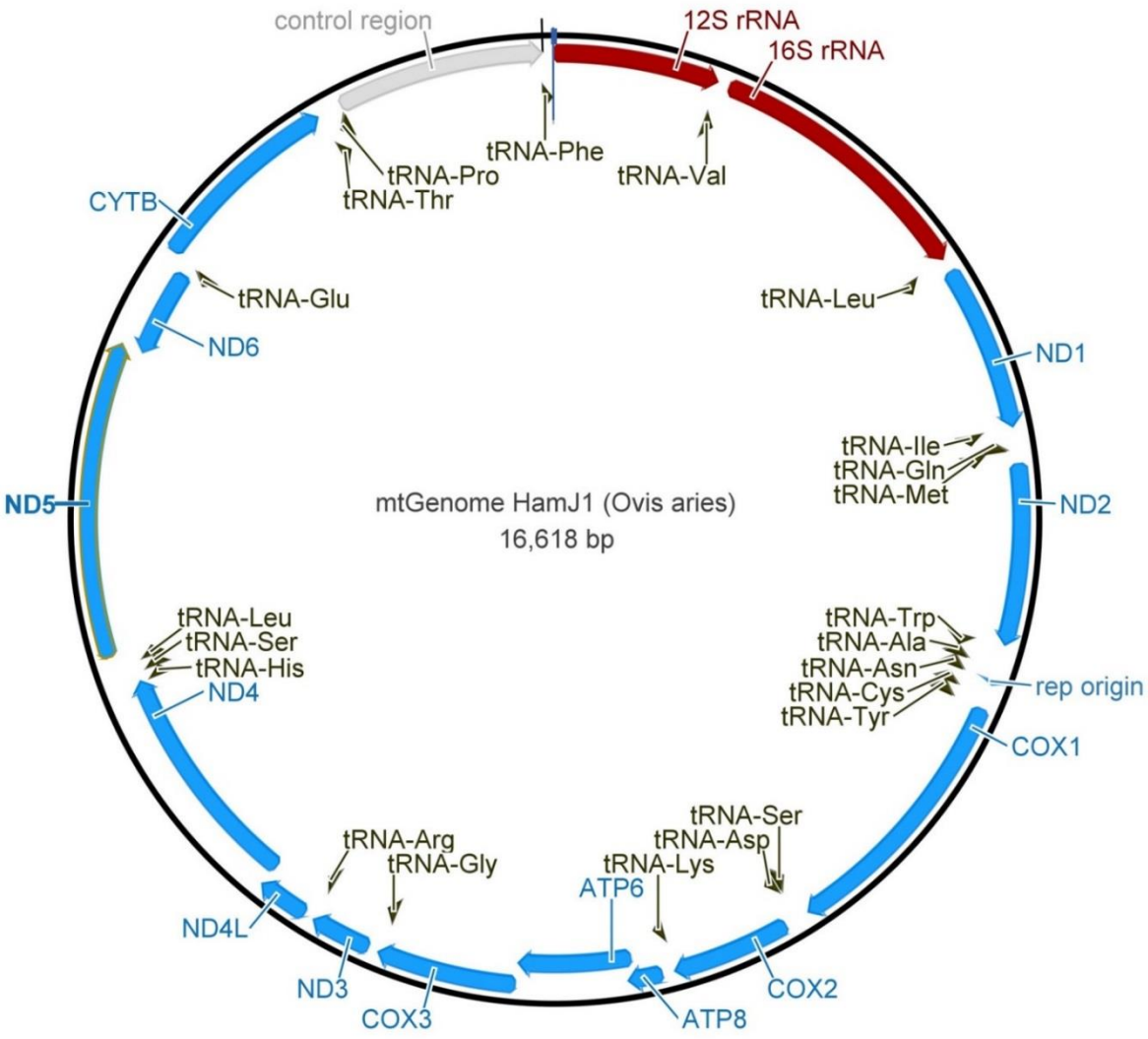

Figure 2. Kurdistani sheep mitogenome map.
The assembled mitogenome HAMJ1 (16,617bp) of *Ovis aries* Hamdani landrace animal Hb4 (GenBank accession number MF004243) with major features: there are 13 protein-coding genes (light blue bars, with the arrow pointing in the transcription directions), 22 tRNA genes (black triangles), the 12S and 16S rRNA genes (dark red) and the D-loop control region (grey). The GC content is 38.9%. For assembly data see Table 1. Equivalent maps of the four other mitogenomes assembled are shown in Supplementary Figures S2-S5

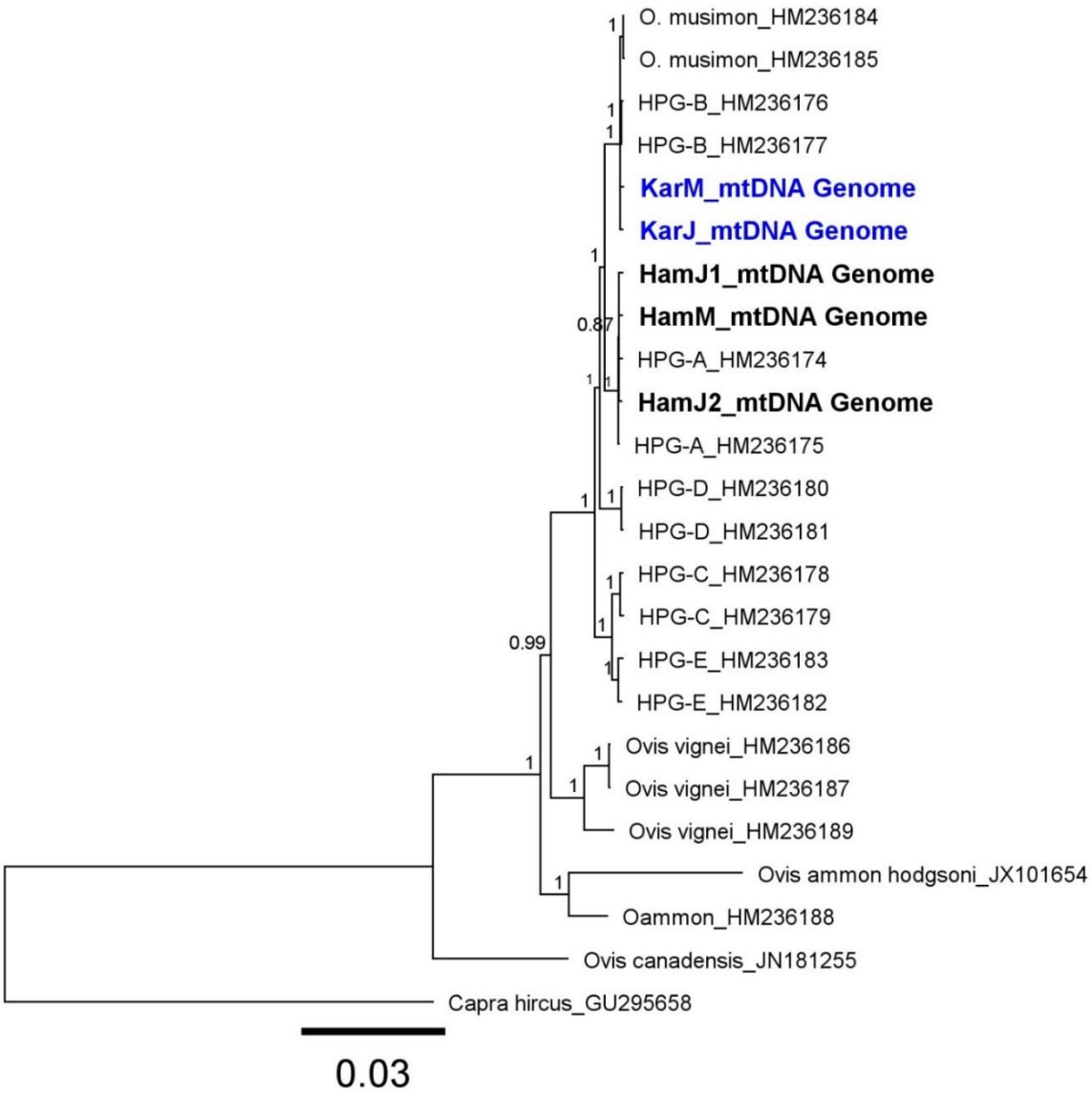

Figure 3. Phylogenetic relationship of sheep mitogenomes.

Bayesian tree showing positions of Hamdani and Karadi mitogenomes in relation to major haplogroups of domestic *O. aries*, two of *O. musimon*, and six wild sheep (*O. ammon*, *O. vignei* and *O. canadensis*). The tree was derived from Bayesian (MrBayes) analysis of alignments of the whole mitochondrial genome sequences. Nodes are labelled with posterior probabilities. The mitochondrial genome of *Capra hircus* was used as outgroup. Supplementary Table S1 gives accession numbers and references for sequences included in the tree.

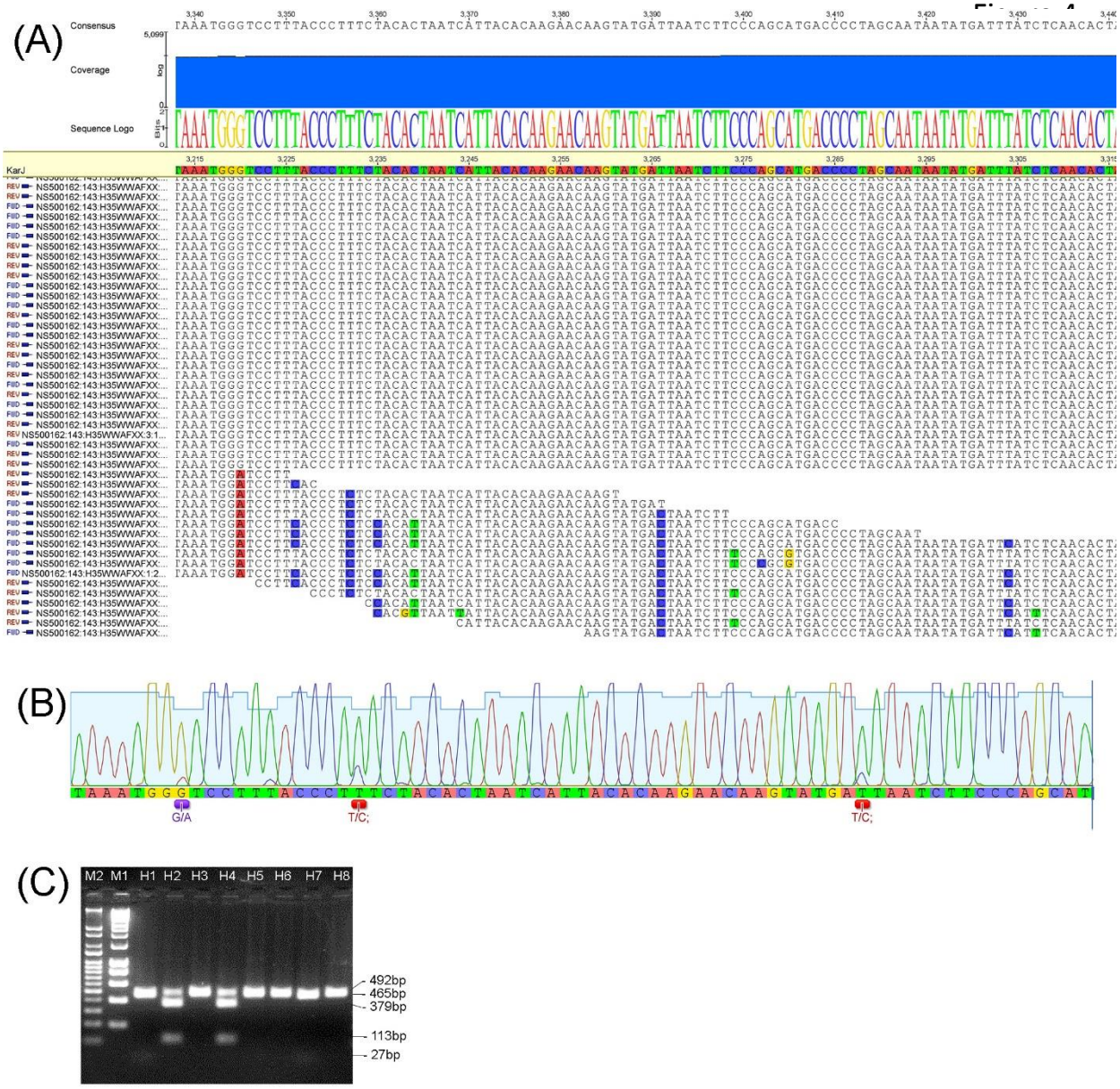

Figure 4. Polymorphisms in mitogenome sequence assembly.
(A) Fragments of some of the 436 150-bp long raw read sequences mapped to a region of the consensus ND1 gene. Seventeen of the reads showed SNPs present in multiple reads (highlighted with colored boxes).
(B) Sanger sequencing trace of PCR products spanning the same region as in (A) showing heterogeneity in base calls (boxes) at the same positions of polymorphisms.
(C) PCR-RFLP patterns of the ND1 mitochondrial gene digested with restriction enzyme AvaII distinguishing three different haplogroups. Haplogroup HPGB with 2 bands (379bp and 113bp) and HPGC with another 2 bands (465bp and 27bp, the latter only visible at longer exposures) and the uncut band represents the haplogroup HPGA. Lanes H2 and H4 show heterogeneity with HPGA as the major haplogroup, and presence of another haplogroup (uncut; 492bp).
Marker lanes are M1: '1kb ladder'; M2:100bp ladder; sample DNA lanes are H1: H-364-P; H2: H-368-P; H3: H-390-P; H4: H-374-P; H5: Hb1-B; H6: H-369-P; H7: Hb2-a; H8: H1a (see Table 2 and Supplementary Table S2).

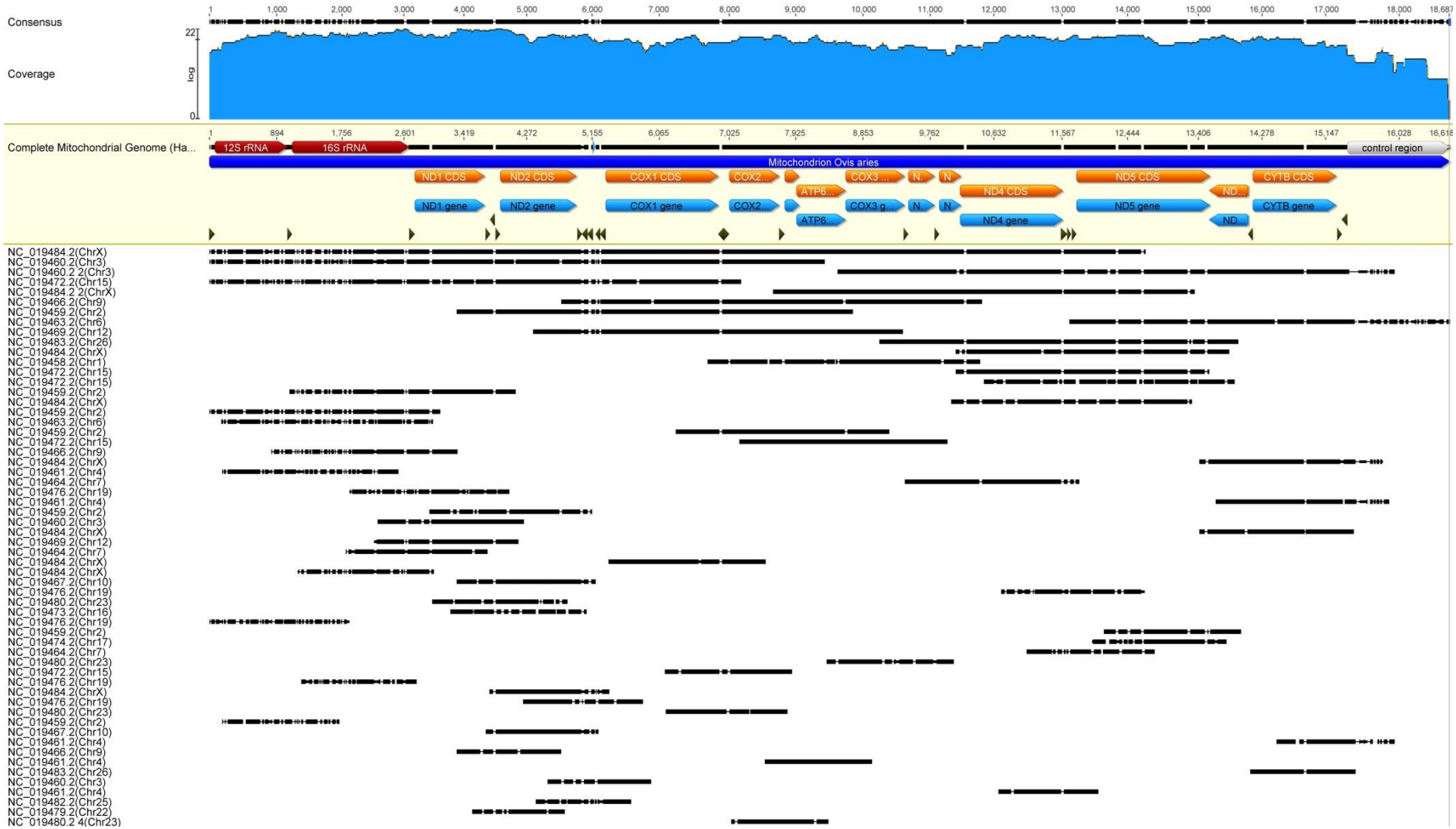

Figure 5. Mitogenome sequences found in nuclear DNA assemblies.

Alignment of selected fragments from chromosome scaffolds in the *Ovis aries* whole genome assembly Oar v4.0 to the sheep mitogenome. Some mitochondrial fragments are present in the nuclear assembly, but would not account for the strong *in situ* hybridization signals seen (Figure 6).

Figure 6

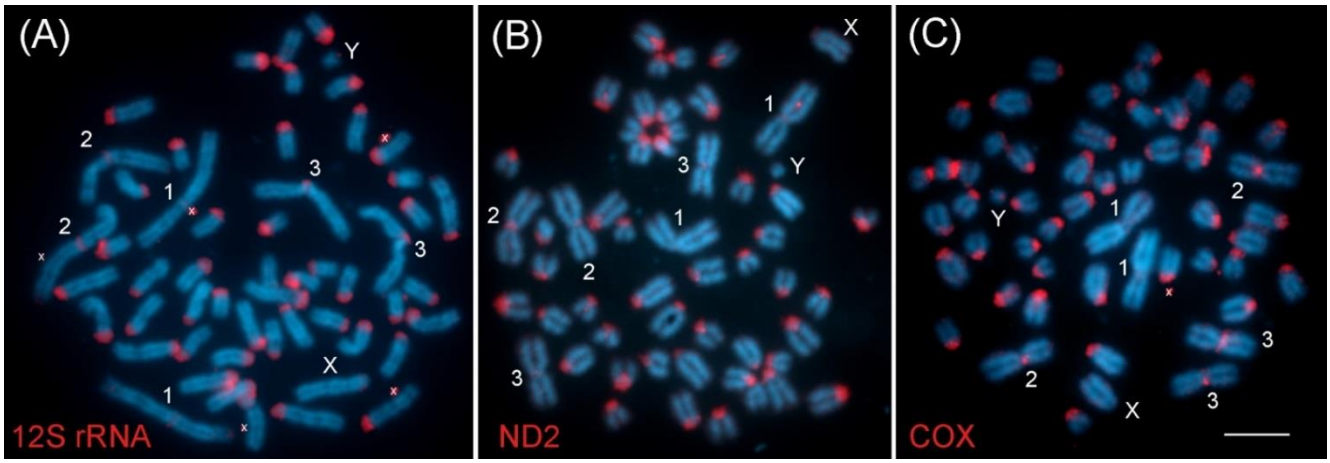

Figure 6. Mitogenome sequences are detected on chromosomes.

Fluorescent *in situ* hybridization of probes (detected in red) to metaphase chromosomes of sheep (2n=54; stained blue with DAPI). Probes were amplified by PCR primers spanning domains of the mitochondrial sequences (see Supplementary Table S2).

(A) 12S rRNA probe; (B) ND2 probe; (C) COX2 probe. All probes give strong signal at the centromeres of all 23 pairs of autosomal acrocentric chromosomes. Hybridization strength to the centromeres of the three pairs of metacentric chromosomes and the X and Y-chromosomes, always weaker, differs between probes (chromosome identifications indicated by numbers and letters). Supplementary Figure S16 shows hybridization results from four additional mitochondrial probes.
Bar = 10µm