ISSN 1022-7954, Russian Journal of Genetics, 2019, Vol. 55, No. 12, pp. 1499–1505. © Pleiades Publishing, Inc., 2019.
Russian Text © The Author(s), 2019, published in Genetika, 2019, Vol. 55, No. 12, pp. 1410–1416.

Targeted Sequencing for Studying Economically Useful Traits and Phylogenetic Diversity of Ancient Sheep[1]

A. A. Kechin[a, b, c, *], M. A. Dymova[a], A. A. Tishkin[d], S. P. Grushin[d], P. K. Dashkovskiy[e], and M. L. Filipenko[a, c, d]

[a] Institute of Chemical Biology and Fundamental Medicine, Siberian Branch of the Russian Academy of Sciences, Novosibirsk, 630090 Russia

[b] Department of Molecular Biology and Biotechnology, Novosibirsk State University, Novosibirsk, 630090 Russia

[c] Department of Clinical Biochemistry, Novosibirsk State University, Novosibirsk, 630090 Russia

[d] Department of Archeology, Ethnography, and Museology, Altai State University, Barnaul, 656049 Russia

[e] Department of Political History, National and State Confessional Relations, Altai State University, Barnaul, 656049 Russia

**e-mail: aa_kechin@niboch.nsc.ru*

Received January 31, 2019; revised April 18, 2019; accepted April 23, 2019

Abstract—Sheep were one of the first animals to be domesticated. The history of sheep domestication and their widespread distribution dates to about ten thousand years ago, during which sheep exhibit both physical changes and modifications at the genetic level. The authors developed a system of 49 oligonucleotide primers for targeted Next Generation Sequencing (NGS) of genetic loci for phylogenetic analysis and identifying economically useful traits. Altogether, NGS libraries were prepared and sequenced on an Illumina MiSeq platform(Illumina) for 48 samples, for 40 of which it was possible to determine phylogenetic lineages: 28 belonged to haplogroup B, 10 to haplogroup A, and one sample each to haplogroups C and D. Study of the genes associated with economically useful traits revealed the samples with nucleotide substitutions in the *MC1R* gene leading to black coat color: two samples with c.218T>A, one with c.361G>A, and two with both substitutions simultaneously, as well as one sample with the substitution in the *GDF8* gene associated with muscle hypertrophy and one with the substitution in the *TYRP1* gene associated with brown coat color. The data obtained confirm a high genetic diversity of sheep from ancient southwestern Siberia and the utility of targeted sequencing for the study of ancient DNA samples.

Keywords: mitochondrial DNA, ancient DNA, archaeology of Siberia, D-loop, SNP, economically useful traits, targeted sequencing, NGS

DOI: 10.1134/S102279541912007X

INTRODUCTION

The domestication of animals reflects a key process in human history, leading to the emergence of new food producing strategies and social organization [1]. The consistent and continuous selection of animals for preferred traits made it possible to obtain less aggressive, but more productive, hardy, and/or quick individuals. The fertility, color, and adaptability to special living conditions could have become other useful properties in domesticated animals.

Sheep (*Ovis aries*) were one of the first animals to be domesticated about 11,000 to 9,000 years ago, most likely, occurring in the area called the Fertile Crescent (the region from North Zagros Mountains to Southeastern Anatolia) [2]. It is considered that sheep selection was initially carried out with respect to meat and milk productivity and only later with respect to wool quality [3]. The genetic determinants, which were hidden for that time, by which selection of these animals was carried out have not only fundamental significance for sheep physiology but also practical significance for people. Understanding the development of economically useful traits in ancient sheep breeds provides a basis for subsequent improvement for modern sheep husbandry.

The history of the emergence and formation of modern breeds can be partially resolved by studying mitochondrial DNA (mtDNA), which reflects maternal inheritance. To date, five maternal lines or haplogroups are distinguished (A, B, C, D, and E), of which the last two are the rarest [4].

While Sanger sequencing is the main sequencing method for both mtDNA and loci associated with economically useful traits, whole-genome sequencing approaches based on massively parallel sequencing

[1] Supplementary materials are available for this article at DOI:@ and are accessible for authorized users.

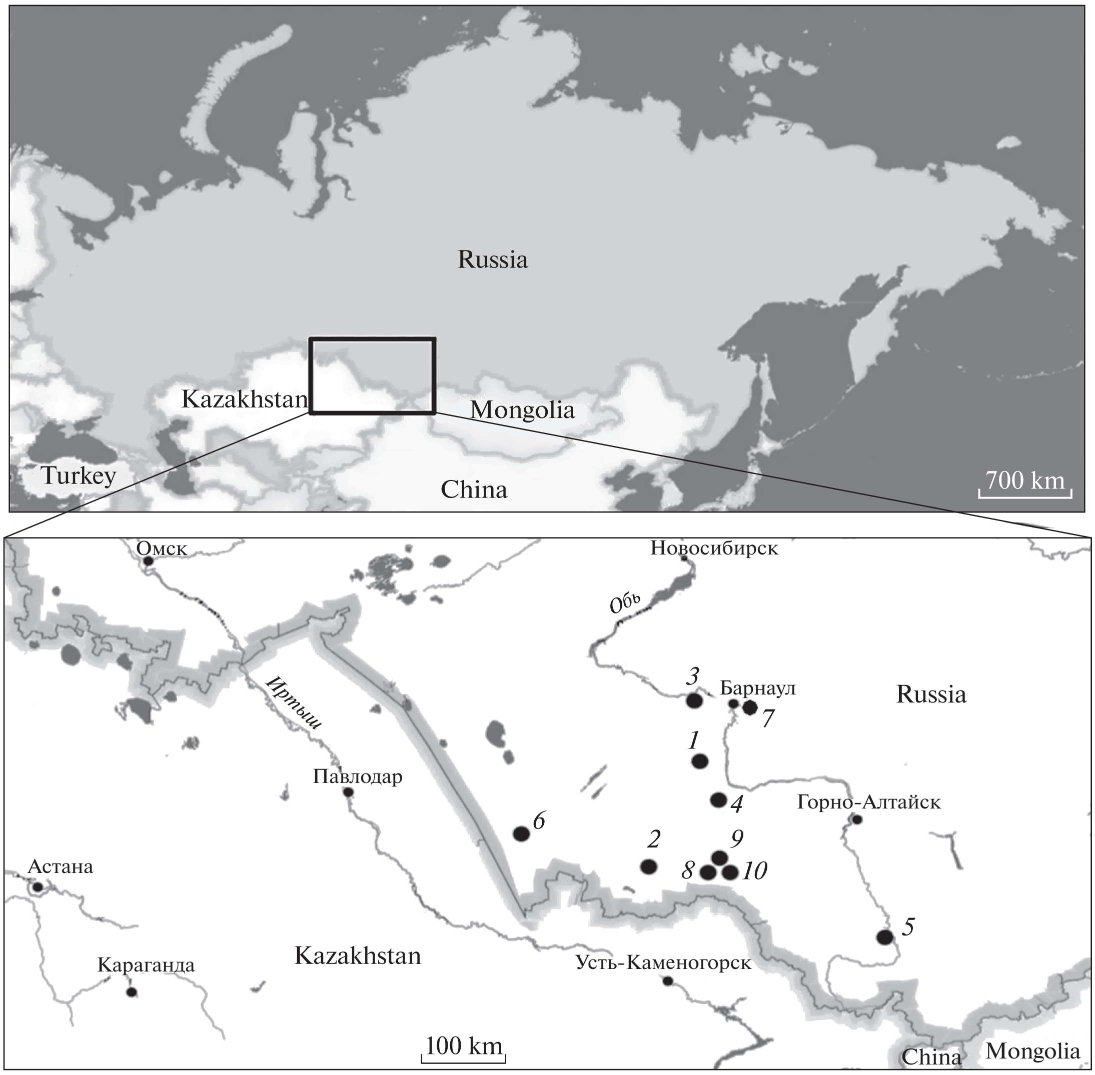

Fig. 1. Schematic map of location of ancient settlements and burial grounds from which the studied samples occur: (*1*) Berezovaya Luka; (*2*) Kolyvanskoe-I; (*3*) Teleutskii Vzvoz-I; (*4*) Myshinyi Log; (*5*) Yaloman II; (*6*) Rublevo-VI; (*7*) Firsovo-XIV; (*8*) Khankarinskii Dol; (*9*) Chineta-II; (*10*) Inskoi Dol.

(MPS or Next Generation Sequencing—NGS) is less often used. On one hand, the latter approach determines the genetic sequences of all required loci, providing the researcher with the most comprehensive genetic information; on the other hand, a significant part of the recovered sequences using NGS is represented by "trash" information, the presence of which does not allow one to make any additional conclusions from the samples.

In this work, the approach that is widely used in clinical applications is targeted sequencing, in which only target genome sequences of the studied object are enriched [5]. The enrichment can be done in several ways: using hybridization and amplification. This is reflected in this study. The applicability of targeted sequencing when studying ancient DNA (**aDNA**) samples was shown to obtain sequences of the markers informative for phylogenetic analysis of ancient sheep populations and to determine genetic determinants of economically useful traits.

MATERIALS AND METHODS

Bone Samples and aDNA Isolation

A detailed description of the samples involved in the work (48 in total) is given in Table 1 of the Supplement. A schematic map of the location of settlements and burial grounds is presented in Fig. 1. aDNA isolation from 0.5 g of ground bone was conducted according to previously described methods [6]. Negative controls were simultaneously used.

Selection of Target Loci and Primer Design

Loci for amplification and subsequent sequencing were selected on the basis of the published data (Table 2 of Supplement). The primers for amplification enrichment of target sequences were automatically constructed by the hi-plex program modified by us [7]. All 1 constructed primers were tested on untargeted hybridization with genome regions (Oar_v4.0 assembly) using the BWA program and Python scripts.

Preparation of NGS Libraries and Sequencing

The preparation of libraries was performed according to the method described previously [5] and including two stages of amplification: production of target sequences and inclusion of adapter and indexing sequences. Sequencing was conducted on the Illumina MiSeq platform using MiSeq reagent kit v3 reagents (150 cycles) according to the manufacturer's instructions.

Analysis of NGS Data

Adapter sequences were removed from the reads by the Trimmomatic program [8]. Reads shorter than 50 nucleotides were not used for further analysis, since the expected length of PCR products together with the primers was about 95 bp. The remaining reads were mapped by the BWA program [9] to the fragments extracted from the genome that correspond to amplified regions. At the same time, closely located fragments (up to 1000 bp) were combined into one. The obtained SAM files were converted into BAM files, reads were sorted, and information about the groups of reads was added using the Picard program (http://broadinstitute.github.io/picard/). The primer sequences were further removed from the reads in 2 BAM files using the cutPrimers program [10]. The genotypes for loci of interest were determined using the samtools program; consensus sequences were obtained using the Pisces program (https://github.com/Illumina/Pisces) and our own Python scripts. Previously, reference sequences [11] were used to determine the mitochondrial haplogroups of the studied samples.

RESULTS AND DISCUSSION

General Statistics of Targeted NGS

To amplify the selected fragments, 49 primer pairs were constructed (Table 3 of Supplement). In total, 22.5 million read pairs were obtained, of which 19 million pairs belonged to the studied 48 samples. There was a median of 386,860 (ranging from 19 to 1,478,630) read pairs per sample. The median percentage of reads that were mapped on the target sequences was 2.31% (ranging from 0.08% to 57.97%). Such low percentages are associated with damaged aDNA, because hybridization of the primers with themselves is more efficient than with DNA matrix. This is also confirmed by the fact that the two largest values (9.30% and 57.97%) were from samples of Mongolian and modern times, respectively.

Phylogenetic Analysis of mtDNA D-Loop Sequences

To determine the maternal lineages (A, B, C, D, and E) of the studied samples, the consensus sequences of D-loop fragments were made (MT: 14652–16501). Each consensus sequence was aligned to reference sequences, the alignments were truncated by the shortest common region, and a phylogenetic tree was constructed by the maximum likelihood algorithm using the MEGA program (version 6.06) [12]. An example of such a tree is shown in Fig. 2. The above D-loop fragment was not recovered for eight samples. Lineage B was the most common in our sample set (28 samples, 70%), and lineage A followed (10, 25%). One sample was found for lineage C (2.5%) and D (2.5%).

The distribution of mitochondrial lineages by geographic location of the sheep samples and chronology are presented in Figs. 3 and 4. As well as in the case of general statistics, most samples found in the Berezovaya Luka settlement and Teleutskii Vzvoz-I burial and memorial complex (memorials with the largest number of samples) belonged to lineages A and B. For the Berezovaya Luka memorial, 83% of samples belonged to lineage B; for the Teleutskii Vzvoz-I, 54%. By dating, the group of samples from these archaeological sites dated to the last quarter of the third to the first quarter of the second millennium BC was the most represented. In it, 72% of samples belonged to lineage B; 24% to lineage A; 3% to lineage C. Thus, we found that lineage B was predominant among all haplogroups represented in our sample (both in individual regions and in different chronological segments).

Genetic Loci Associated with Economically Useful Traits

When sequencing the Illumina libraries, 78% of target regions were read. For most loci, the samples were homozygous for wild type alleles. Five samples from the settlement of the Early Bronze Age, Berezovaya Luka, were heterozygous for one or several single nucleotide substitutions: two for *MC1R* c.218T>A, one for *MC1R* c.361G>A, and two for both loci. The *MC1R* gene is associated with black wool color with an autosomal dominant type. One sample from Teleutskii Vzvoz-I was heterozygous for the locus *GDF8* g.118146642 G>A; one sample from the settlement Berezovaya Luka was heterozygous for *TYRP1* c.869 G>T. The first locus is associated with muscle hypertrophy (an increase in the muscle mass) [13]; the second is associated with brown wool color, and an autosomal dominant inheritance of the trait was demonstrated

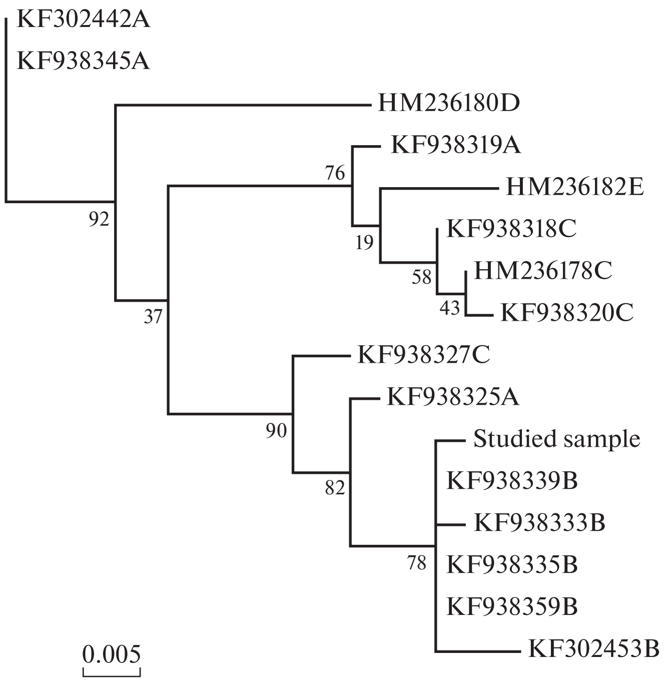

Fig. 2. Example of phylogenetic tree according to which belonging to a phylogenetic line was determined (the last letter in the names of reference samples indicates the line; numbers near the nodes reflect the support level of the corresponding node).

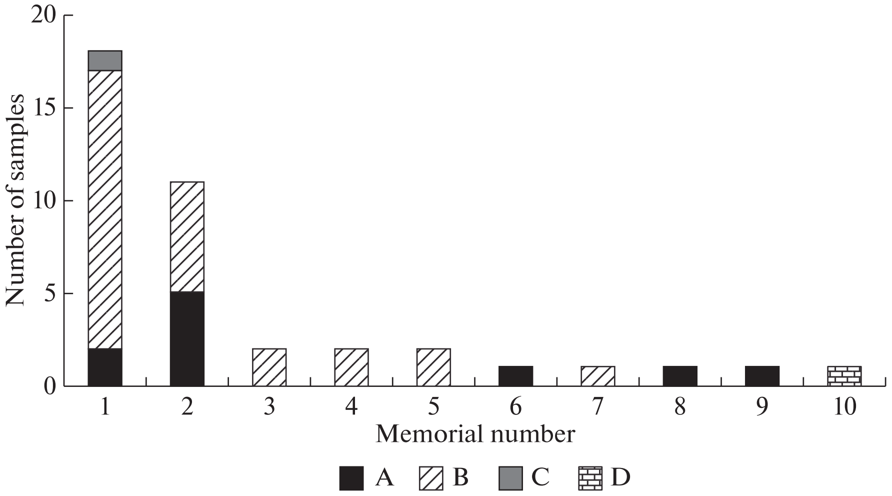

Fig. 3. Distribution of samples by phylogenetic lines in places of their discovery. (**1**) Berezovaya Luka; (**2**) Teleutskii Vzvoz-I; (**3**) Chineta-II; (**4**) Kolyvanskoe-I; (**5**) Khankarinskii dol; (**6**) Firsovo-XIV; (**7**) Yaloman-II; (**8**) Myshinyi Log-I; (**9**) Rublevo-VI; (**10**) Central Mongolia.

for the second locus [14]. The detected variants were confirmed by Sanger sequencing (Fig. 5).

Thus, among the studied 48 sheep samples, five exhibited genetic determinants for black wool color, one exhibited genetic determinants for brown wool, and one exhibited genetic determinants for muscle hypertrophy.

Examining genetic variation of ancient animal remains is essential for resolving processes of evolution and domestication. Whole genome sequencing is one of the most common and also most simple and convenient methods of genetic analysis of ancient DNA. At the same time, a misuse of massively parallel sequencing power on the genome regions, reading of which

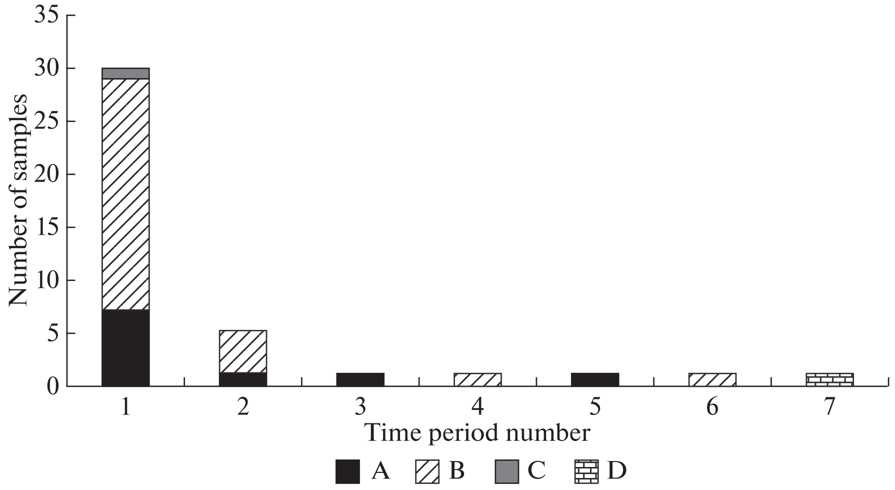

Fig. 4. Distribution of samples by phylogenetic lines and dating: (**1**) last quarter of third–first quarter of second millennium BC, Early Bronze Age; (**2**) fifth–third centuries BC, Early Iron Age; (**3**) tenth–eighth centuries BC, Late Bronze Age; (**4**) second half of fourth–first half of fifth centuries AD (Zhuzhan time), Early Iron Age; (**5**) ninth–tenth centuries AD, Early Middle Ages; (**6**) 13th–14th centuries AD, Mongolian time; (**7**) 21st century AD, modern period.

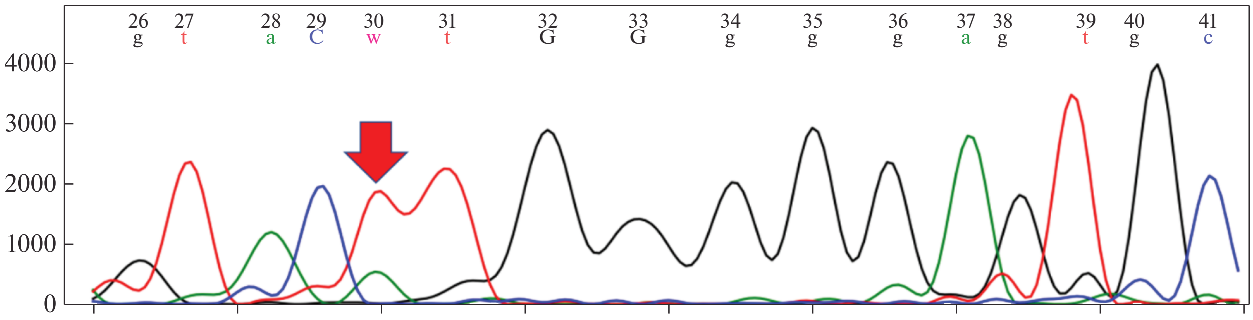

Fig. 5. Example of sequenogram to confirm the presence of variation in the studied samples. Heterozygous A>T substitution, which corresponds to c.218T>A substitution in the *MC1R* gene, is marked by an arrow.

does not carry any information useful for the study, is marked as its significant disadvantage. In this regard, we developed an amplification panel for targeted sequencing of sheep genomic loci associated with informative phylogenetic variation and economically useful traits (Table 2 of Supplement). Previously, attempts to study such systems for ancient DNA samples were made [15]. However, they are widely used only in clinical studies of DNA samples isolated from paraffin-embedded histological blocks and blood plasma [16], which is associated with the complexity of conducting multiplex amplification of ancient DNA. At the same time, targeted sequencing makes it possible to obtain genetic sequences of many regions in a single run, which significantly reduces the costs of recovering desired genetic material.

Using the developed NGS panel, it was possible to obtain a rather high coverage of most selected loci (78%). No specific reads were developed for the remaining loci. This was probably associated with the structure of the appropriate primers, for which the formation of secondary structures was thermodynamically more advantageous. Together, the percentage of reads that were mapped to the target sequences was much higher for more modern samples (58.0 and 9.3% against median 2.3% for others).

Determination of phylogenetic lineages according to nucleotide sequences of the mitochondrial D-loop region was one of the main results of the work done. Most of the samples were assigned to lineage B (70% of samples) and A (25%). Lineages C and D accounted for only 2.5% each, and haplogroup E was determined only for one sheep sample from a modern animal. The obtained distribution of maternal lineages corresponds to those described in the literature [17]. It is possible that the predominance of lineages A and B in the studied region is associated with their first appearance here as compared with other lineages [18]. However, other explanations are also possible. It also remains unclear why lineages A and B were first present in southwestern Siberia, while haplogroup C was poorly represented.

We detected samples that exhibited genetic determinants of black and brown wool color. There is an important fact that, along with the samples heterozygous for one of the loci, a sample heterozygous for two *MC1R* gene loci (c.218T>A and c.361G>A) was detected, from which it follows that individuals carrying one homozygous or heterozygous substitution in each of these loci coexisted in the same sheep population. This suggests high genetic diversity of sheep of southwestern Siberia at the end of the third and beginning of the second millennium BC.

Thus, in order to increase the efficiency of using NGS when studying aDNA, we developed a method for preparing target NGS libraries which made it possible to conduct phylogenetic analysis and to obtain nuclear sequences associated with economically useful traits. This method can be used both for studying the same loci in other samples and as a basis for the development of new NGS panels, by which other loci of interest in sheep and other animals can be studied. Higher coverage of the target regions is a main advantage with the whole genome sequencing along with lower costs of extracted genetic material, compared to Sanger sequencing. The results of applying the described approach enhance our knowledge about evolution and initial spread of domesticated sheep to southwestern Siberia.

ACKNOWLEDGMENTS

We are grateful to D.V. Panin and Ya.V. Frolov for the samples from archaeological sites Rublevo-VI and Firsovo-XIV provided for the comparative analysis and to the native English speaker T.R. Hermes for thorough text revision.

FUNDING

This work was supported by the Russian Science Foundation (project no. 16-18-10033 “Formation and Evolution of Life Support Systems in Nomadic Societies of Altai and Adjacent Territories in the Late Antiquity and Middle Ages: Complex Reconstruction”).

COMPLIANCE WITH ETHICAL STANDARDS

Conflicts of interest. The authors declare that they have no conflict of interest.

Statement on the welfare of animals. This article does not contain any studies involving animals performed by any of the authors.

Statement of compliance with standards of research involving humans as subjects. This article does not contain any studies involving human participants performed by any of the authors.

REFERENCES

1. Vigne, J.D., The origins of animal domestication and husbandry: a major change in the history of humanity and the biosphere, *C. R. Biol.*, 2011, vol. 334, no. 3, pp. 171–181. https://doi.org/10.1016/j.crvi.2010.12.009
2. Zeder, M.A., Domestication and early agriculture in the Mediterranean Basin: origins, diffusion, and impact, *Proc. Natl. Acad. Sci. U.S.A.*, 2008, vol. 105, no. 33, pp. 11597–11604. https://doi.org/10.1073/pnas.0801317105
3. Ryder, M., *Sheep and Man*, Duckworth, 1983.
4. Meadows, J.R.S., Hiendleder, S., and Kijas, J.W., Haplogroup relationships between domestic and wild sheep resolved using a mitogenome panel, *Heredity* (Edinburgh), 2011, vol. 106, no. 4, pp. 700–706. https://doi.org/10.1038/hdy.2010.122
5. Ermolenko, N.A., Boyarskikh, U.A., Kechin, A.A., et al., Massive parallel sequencing for diagnostic genetic testing of BRCA genes—a single center experience, *Asian Pac. J. Cancer Prev.*, 2015, vol. 16, no. 17, pp. 7935–7941. https://doi.org/10.7314/apjcp.2015.16.17.7935
6. Pääbo, S., Gifford, J.A., and Wilson, A.C., Mitochondrial DNA sequences from a 7000-year old brain, *Nucleic Acids Res.*, 1988, vol. 16, no. 20, p. 9775. https://doi.org/10.1093/nar/16.20.9775
7. Nguyen-Dumont, T., Pope, B.J., Hammet, F., et al., A high-plex PCR approach for massively parallel sequencing, *Biotechniques*, 2013, vol. 55, no. 2, pp. 69–74. https://doi.org/10.2144/000114052
8. Bolger, A.M., Lohse, M., and Usadel, B., Trimmomatic: a flexible trimmer for Illumina sequence data, *Bioinformatics*, 2014, vol. 30, no. 15, pp. 2114–2120. https://doi.org/10.1093/bioinformatics/btu170
9. Li, H. and Durbin, R., Fast and accurate short read alignment with Burrows–Wheeler transform, *Bioinformatics*, 2009, vol. 25, no. 14, pp. 1754–1760. https://doi.org/10.1093/bioinformatics/btp324
10. Kechin, A., Boyarskikh, U., Kel, A., and Filipenko, M., 2 cutPrimers: a new tool for accurate cutting of primers from reads of targeted next generation sequencing, *J. Comput. Biol.*, 2017, vol. 24, no. 11, pp. 1138–1143. https://doi.org/10.1089/cmb.2017.0096
11. Dymova, M.A., Zadorozhny, A.V., Mishukova, O.V., et al., Mitochondrial DNA analysis of ancient sheep from Altai, *Anim. Genet.*, 2017, vol. 48, no. 5, pp. 615–618. https://doi.org/10.1111/age.12569
12. Tamura, K., Stecher, G., Peterson, D., et al., MEGA6: molecular evolutionary genetics analysis version 6.0, *Mol. Biol. Evol.*, 2013, vol. 30, no. 12, pp. 2725–2729. https://doi.org/10.1093/molbev/mst197
13. Clop, A., Marcq, F., Takeda, H., et al., A mutation creating a potential illegitimate microRNA target site in the myostatin gene affects muscularity in sheep, *Nat. Genet.*, 2006, vol. 38, no. 7, pp. 813–818. https://doi.org/10.1038/ng1810
14. Hinten, G.N., Hale, M.C., Gratten, J., et al., SNP-SCALE: SNP scoring by colour and length exclusion,

Mol. Ecol. Notes, 2007, vol. 7, no. 3, pp. 377—388. https://doi.org/10.1111/j.1471-8286.2006.01648.x

15. Stiller, M., Knapp, M., Stenzel, U., et al., Direct multiplex sequencing (DMPS)—a novel method for targeted high-throughput sequencing of ancient and highly degraded DNA, *Genome Res.*, 2009, vol. 19, no. 10, pp. 1843—1848. https://doi.org/10.1101/gr.095760.109
16. Kechin, A., Khrapov, E., Boyarskikh, U., et al., BRCA-analyzer: automatic workflow for processing NGS reads of BRCA1 and BRCA2 genes, *Comput. Biol. Chem.*, 2018, vol. 77, pp. 297—306. https://doi.org/10.1016/j.compbiolchem.2018.10.012
17. Tapio, M., Marzanov, N., Ozerov, M., et al., Sheep mitochondrial DNA variation in European, Caucasian, and Central Asian areas, *Mol. Biol. Evol.*, 2006, vol. 23, no. 9, pp. 1776—1783. https://doi.org/10.1093/molbev/msl043
18. Lv, F.-H., Peng, W.-F., Yang, J., et al., Mitogenomic meta-analysis identifies two phases of migration in the history of eastern Eurasian sheep, *Mol. Biol. Evol.*, 2015, vol. 32, no. 10, pp. 2515—2533. https://doi.org/10.1093/molbev/msv139

Translated by A. Barkhash

SPELL: 1. untargeted, 2. cutPrimers