The first draft genome assembly of Snow sheep (*Ovis nivicola*)

Maulik Upadhyay¹*, Andreas Hauser²*, Elisabeth Kunz¹*, Stefan Krebs², Helmut Blum², Arsen Dotsev³, Innokentiy Okhlopkov⁵, Vugar Bagirov³, Gottfried Brem⁴, Natalia Zinovieva³, Ivica Medugorac¹#

¹Population Genomics Group, Department of Veterinary Sciences, LMU Munich, Lena-Christ-Str. 48, 82152 Martinsried, Germany

²Laboratory for Functional Genome Analysis, Gene Center, LMU Munich, 80539 Munich, Germany.

³L.K. Ernst Federal Science Center for Animal Husbandry, Moscow Region, Podolsk, Russian Federation

⁴Institute of Animal Breeding and Genetics, University of Veterinary Medicine, Vienna, Austria

⁵Institute for Biological Problems of Cryolithozone, Yakutsk, Russian Federation

*These authors contributed equally to this work.

#corresponding author

Maulik Upadhyay: U.Maulik@gen.vetmed.uni-muenchen.de

Andreas Hauser: Andreas.Hauser@LMU.de

Elisabeth Kunz: Elisabeth.Kunz@gen.vetmed.uni-muenchen.de

Stefan Krebs: krebs@genzentrum.lmu.de

Helmut Blum: blum@genzentrum.lmu.de

Arsen Dotsev: asnd@mail.ru

Innokentiy Okhlopkov: imo-ibpc@yandex.ru

Vugar Bagirov: vugarbagirov@mail.ru

Gottfried Brem: Gottfried.Brem@vetmeduni.ac.at

Natalia Zinovieva: n_zinovieva@mail.ru

Ivica Medugorac: Ivica.Medjugorac@gen.vetmed.uni-muenchen.de
Abstract

The snow sheep, *Ovis nivicola*, which is endemic to the mountain ranges of northeastern Siberia, are well-adapted to the harsh cold climatic conditions of their habitat. In this study, using long reads of Nanopore sequencing technology, whole genome sequencing, assembly and gene annotation of a snow sheep was carried out. Additionally, RNA-seq reads from several tissues were also generated to supplement the gene prediction in Snow sheep genome. The assembled genome was ~2.62 Gb in length and was represented by 7,157 scaffolds with N50 of about 2 Mb. The repetitive sequences comprised of 41% of the total genome. BUSCO analysis revealed that the snow sheep assembly contained full-length or partial fragments of 97% of mammalian universal single-copy orthologs (n=4,104), illustrating the completeness of the assembly. In addition, a total of 20,045 protein coding sequences were identified using comprehensive gene prediction pipeline. Of which 19,240 (~96%) sequences were annotated using protein databases. Moreover, homology-based searches and de-novo identification detected 1,484 tRNAs, 243 rRNAs, 1,931 snRNAs, and 782 miRNAs in the snow sheep genome. To conclude, we generated the first *de novo* genome of the snow sheep using long reads; these data are expected to contribute significantly to our understanding related to evolution and adaptation within the *Ovis* genus.

**Keywords:** Snow sheep, promethION sequencing, de novo assembly, annotation
Introduction

The *Ovis* genus is characterized by its abundance of species and subspecies that inhabit various habitats around the world. The most recognized between scientists is the classification, suggested by Nadler et al. (1973), which distributes the wild *Ovis* in seven species: *Ovis ammon*, *Ovis musimon*, *Ovis orientalis*, *Ovis vignei*, *Ovis dalli*, *Ovis canadensis*, and *Ovis nivicola*. This classification was used by Rezaei H.R. et al. (2010) (Rezaei, et al. 2010), who studied the taxonomy of wild ovis using CytB sequence. However, the International Council for Game and Wildlife Conservation (IC) use the classification as suggested by Valdez R. (1982) (Raul Valdez 1982), who classified wild Ovis in six species. *Ovis nivicola*, also known as snow sheep (Fig. 1A), is endemic to the mountain ranges of northeastern Siberia between the Lena River in the west and the Chukotka and Kamchatka peninsula in the east. A secluded subspecies of snow sheep is also found more west on the Putorana Plateau (Zheleznov-Chukotskii 1994). Due to intensified anthropogenic activities, snow sheep populations have retreated to even higher mountain areas in recent times. The remoteness and harsh climatic conditions of their habitat make these wild sheep the ideal candidates to study mammalian adaptation to extreme environments with regard to elevation and temperature. At the same time, these factors might also be among the reasons why only few genetic studies on snow sheep have been conducted so far. While reference genomes for domestic sheep (Archibald, et al. 2010) as well as the wild sheep species *Ovis canadensis* (Miller, et al. 2015) and *Ovis ammon* (Yang, et al. 2017) have been available for several years now, only a mitochondrial reference genome has been published for snow sheep (Dotsev, et al. 2019). Investigations of the genetics of different snow sheep populations have mainly been carried out on microsatellite (Deniskova, et al. 2018) and SNP genotype level (Deniskova, et al. 2016; Dotsev, et al. 2018; Medvedev, et al. 2017) so far. In the present study, short-read Illumina and long-read Nanopore sequence data was generated from snow sheep samples and used for *de novo* assembly of the snow sheep genome. This assembly will build the basis for whole genome sequence-based studies on the genetic characteristics of snow sheep that could also shed light on their demographic history and adaptation to the extreme cold environments.

Material and Methods

Sample Collection and Processing
For whole genome sequencing using PromethION and Illumina sequencing data, the muscle sample was collected from a healthy male. For RNA sequencing, samples of six different tissues (skin, cerebrum, heart, liver, large intestine) were taken from three different snow sheep from the Suntar-Hayat and Central Verkhoyansk Ridges in Russia. The sampling locations were selected after observations from the air in 2008–2010 and covered an area between 63 and 71 degrees north latitude, and between 128 and 139 degrees east longitude. The samples were placed in liquid nitrogen immediately after the collection. They were later stored at -80°C for later analysis.

**Construction of Sequencing Libraries**

Libraries for long read sequencing were prepared from app. 3 µg of unsheared genomic DNA, following the protocol of Oxford Nanopore’s LSK109 kit (ONT, Oxford, UK). First DNA was end-repaired and A-tailed using the Ultra II endrepair module (New England Biolabs, Ipswich MA, USA) for 10 min at 20 °C. Enzymes were inactivated at 65°C for 5 min and DNA was purified with 1x volumes of Ampure XP beads (Beckman Coulter, Brea CA, USA). End-repaired DNA was eluted from the beads at 55°C for 20 min. Then sequencing adapter with attached motor protein was ligated to the DNA fragments using T4 Quick ligation module (New England Biolabs, Ipswich MA, USA) and components from LSK109 sequencing kit. Adapted library was purified with 0.45 volumes of Ampure beads and eluted in LSK109 EB buffer for 20 min at 37°C. App. 400 ng of library was loaded on a PromethION flowcell and run for 72h on a PromethION beta sequencer (ONT, Oxford, UK).

The Library construction for the Illumina sequencing was carried out with 200 ng of genomic DNA following the Illumina library preparation protocols (NexteraFlex DNA library kit, Illumina, San Diego, USA). The resulting amplified library was quantified and controlled on an Agilent Bioanalyser 2100 (Agilent, Santa Clara USA) and sequenced in 2*100bp paired-end mode on an Illumina HiSeq1500, yielding approximately 200 Mio read pairs.

For RNAseq library construction total RNA was prepared from tissue of hunted animals. Small tissue pieces were collected into tubes containing solution D (Chomczynski and Sacchi 1987). Preserved tissue was stored and transported at -80°C, homogenized in trizol (Silent crusher, Heidolph, Germany) and extracted on a Maxwell RSC48 device using miRNA tissue kit (both Promega, Madison, USA). Prior to constructing RNA-seq libraries, total RNA was
analysed on Bioanalyser (Agilent, Santa Clara, USA) and samples with least degraded RNA (RIN>6) were used for RNAseq library generation with Lexogen sense mRNA (Lexogen Vienna, Austria) and Nugen Complete (Redwood City, USA) RNAseq kits. Libraries were sequenced on Illumina HiSeq1500.

**Genome Assembly**

Porechop version 0.2.4 was used for adapter trimming and Guppy version 3.5.1 was used for base calling. After evaluating a number of established and upcoming long read assemblers, wtdbg2 (redbean) version 2.5 (Ruan and Li 2019) with default settings for Oxford Nanopore reads (option -x ont) but with shorter acceptable reads (-L 2000) was chosen based on giving by far the most contiguous assemblies. The assemblies generated from the base-called reads using different Guppy versions were polished with up to three rounds of Racon version 1.3.3. (Vaser, et al. 2017), from which consistently round 2 was chosen based on having the highest count of Illumina reads mapping to it, which implied that round 3 is overcorrecting. The Illumina reads were further used for one round of polishing with Pilon version 1.23 (Walker, et al. 2014).

**Annotation of Repeat Sequences**

RepeatModeler version 1.0.11 (Smit and Hubley 2008-2015) was used to identify a de novo repeat genomic sequences in the snow sheep genome. Subsequently, this custom repeat library was used in RepeatMasker version 4.0.9 (Smit, et al. 2013-2015; Tarailo-Graovac and Chen 2004) to soft-mask the identified repeat families. Additionally, the trf tool version 4.07b (Benson 1999) was employed to predict the tandem repeats.

**Scaffolding Using RNA-Seq Data**

To further improve the assembly, RNA-seq based approach as implemented in AGOUTI version 0.3.3 was used (Zhang, et al. 2016). For this purpose, the paired-end RNA-seq data sets from five tissues were aligned to the assembled snow sheep genome using Hisat2 version 2.1.0 (Kim, et al. 2015) aligner with default parameters. Augustus version 3.3.3 (Stanke and Waack 2003) was used to predict gene models from the snow sheep assembly using configuration parameters trained on Homo sapiens (human). The name-sorted bam file and gene models were used as inputs in AGOUTI (Zhang, et al. 2016) to link the contigs.

**Genome Annotation**
Three different approaches, namely, *ab initio* prediction, protein-homology, and RNA-seq based annotation were carried out to predict the gene structure. For protein-homology based approaches, the reference protein sequences from NCBI database of goat (*Ovis aries*) and cattle (*Bos taurus*) were downloaded. Additionally, proteins sequences of ncbi refseq database of Domestic sheep (*Oar_rambouillet_v1.0*) and reviewed uniprot sequences of mammals were also downloaded. All these proteins sequences were aligned against the snow sheep genome using Spaln2 version 2.4.0 (Iwata and Gotoh 2012). For RNA-seq based annotation, the RNA-seq reads of five different tissues (cerebrum, heart, large intestine, liver and skin), which consisted of paired-end as well as single-end reads, were filtering using *trim_galore* version 1.4 (Martin 2011). Subsequently, all the filtered reads were *de novo* assembled using *Trinity* version 2.10.0 (Grabherr, et al. 2011) (parameters: “-no_normalize_reads --min_kmer_cov 1 -SS_lib_type F). To reduce the redundancy of the assembled transcripts, we used the “tr2aacds” tool as implemented in the EvidentialGene (Gilbert 2019) pipeline. In the next step, Program to Assemble Spliced Alignments (PASA) version 2.0.0 (Haas, et al. 2003) was used to obtain the refined gene models from the assembled transcripts. The *ab initio* prediction of gene structure was carried out SNAP version 2006-07-28 (Korf 2004), Augustus version 3.3.3 (Stanke, et al. 2006) and Geneld version 1.4 (Parra, et al. 2000) using the parameters either trained for human or mammals. Hisat2 with default parameters was used to align the RNA-seq reads against the snow sheep genome. Later, these aligned reads were assembled into gene models using *Stringtie* version 1.3.6 (Pertea, et al. 2015). Subsequently, gene-models predicted by all the above approaches were provided as inputs for the EvidenceModeler (EVM) version 1.1.1 tool (Haas, et al. 2008) to generate the non-redundant sets of gene structure. The final set of high-quality protein-coding sequences were also prepared based on the following criteria: 1). A gene should be supported by all the three methods of *ab-initio* predictions, or 2). A gene should at least be supported either by transcript-based evidence or protein-homology based evidence. The amino-acid sequences were functionally annotated using *emapper*-2.0.1.4 (Huerta-Cepas, et al. 2017) based on eggNOG orthology data (Huerta-Cepas, et al. 2018). For this purpose, sequence searchers were performed using DIAMOND (Buchfink, et al. 2015). Additionally, functional annotation of protein sequences was also carried out using InterProScan 5.36.75.0 database (Jones, et al. 2014). Both these procedures also assigned the gene ontology terms associated with the protein function.

**Identification of Non-coding RNAs (ncRNAs)**
Various non-coding RNAs (ncRNAs) were also predicted using combination of different tools and Rfam database. To identify, cytoplasmic transfer (t)RNA gene, tRNAscan-SE version 2.0.5 (Chan, et al. 2019) was used with the default settings. Further, tRNAs were filtered based on the following criteria: 1). It overlapped with the SINEs identified by Repeatmasker, 2). It was identified as pseudogene, 3). It had mismatched isotypes. Further, miRNAs, small nuclear (SnRNAs), nucleolar (SnoRNAs), and rRNA were annotated by searching the Rfam database (Griffiths-Jones, et al. 2005) (release 14.1) with Infernal version 1.1.3 (Nawrocki and Eddy 2013). We also annotated ribosomal RNA genes using RNAmer version 1.2 (Lagesen, et al. 2007). Subsequently, we filtered out rRNAs identified using Rfam database that had no overlap with rRNAs identified using RNAmer.

**Result and Discussion**

**Genome Assembly and Quality Assessment**

A total of 56 Gb data in 17 million PromethION sequences (Fig. 1B) with a read N50 of 7,029 was used to assemble the snow sheep genome. The genome assembly (Fig. 1C) resulted in 7,373 contigs and was estimated to be 2.62 Gb in length with L50 (Length at N50) and L90 (Length at N90) of 1.76 Mb (N50=438) and 323.32 Kb (N90=1,674), respectively. The size of the assembly comparable to domestic sheep (*Ovis aries, Oar 4.1, ~2.61 Gb*) but it is smaller than goat genome assembly (*Capra hircus, ARS1, 2.92 Gb*). The completeness of the assembly was assessed using BUSCO tool (v 3.1.0); the single copy orthologs set in mammalian lineage were searched against the assembled genome of snow sheep using BUSCO tool (v 3.1.0). The results (Table S1) indicated that the assembly covered 3,947 (~97%) of the total 4,104 orthologs. Of these 3,947 genes, 3,762 (~92%) were completely covered in the assembled genome, while only about 3% were assumed to be missing in the assembly. Additionally, the assessment of the base content also indicated that GC content of the assembly was 42.12% which was comparable to that of domestic sheep (41.9%) and goat (41.5%).

**Repeat Annotation**

RepeatModeler identified 605 repeat consensus sequences in Snow sheep genome. Among these, LINE (long interspersed nuclear elements) had highest number of consensus followed by LTR (long terminal repeats) consensus. Subsequently, RepeatMasker procedure (Table S2) soft-masked about 1,087 Mb (~41.5%) sequences of the genome. An interspersed repeat...
landscape, to study the divergence of transposable element classes, was produced for the snow sheep assembly using the scripts calcDivergenceFromAlign.pl and createRepeatLandscape.pl as provided with RepeatMasker package. The resulting landscape (Fig. 2A) identified LINE repeat families as the most abundant, constituting about 30% of the snow sheep genome (Table S2). The lowest substitution level in some LINE L1 copies suggested that these are the youngest repeat elements and are still probably expanding and diversifying in snow sheep genome.

**RNA-based Scaffolding and Quality Assessment**

RNA-based scaffolding approach (Fig. 1C) reduced the number of contigs from 7,373 to 7,157 and it also increased L50 from ~1.76 Mb to ~2.01 Mb, and L90 from ~323 Kb to ~346 Kb. After carrying out RNA-based scaffolding, we aligned the Illumina paired-end sequences, of the same individual from which the assembly was generated, to the assembly using bwa-mem and the result suggested that about 99.42% reads were mapped and 97.6% were aligned as proper pairs. In another approach, to assess the completeness of the snow sheep assembly, we aligned the snow sheep assembly against the latest domestic sheep assembly (**Oar_rambouillet_v1.0**) available in NCBI genome database. The results (Fig. 2B) suggested that scaffolds of the snow sheep assembly covered the entire genome of domestic sheep assembly.

**Genome Annotation and Quality Assessment**

A total of 57,267 protein coding sequences (Table S3) were identified using comprehensive gene prediction pipeline. After filtering (refer to “Material and Method” section for detail) these were reduced to 20,202 high-quality protein sequences. Of these 20,202, 18,870 proteins were annotated using InterPro database, while 19,175 protein sequences were annotated using eggNOG orthology data. Of these 19,175 protein sequences, 18,919 proteins were mapped to mammalian orthologous gene families. Combining the results of **InterProScan** and **emapper**, a total of 19,240 (~96%) protein sequences were annotated. To assess the quality of protein-coding gene set of the snow sheep assembly, BUSCO was run in the protein mode against the mammalian lineage. The results indicated that the annotation successfully captured about ~81% complete and ~14% fragmented BUSCOs, and only ~5% of BUSCOs were missing from the predicted protein sequences. Moreover, DOGMA version 3.0 (Dohmen, et al. 2016) was also used to measure the completeness of a proteome. It measures the
completeness of conserved protein domains identified in a given proteome by providing it as a percentage of a defined core set. The DOGMA analysis based on the domain annotation by PfamScan annotation revealed that about 90% of the total expected conserved domain arrangements were present in the proteome of snow sheep assembly.

Annotation of ncRNAs

A total of 1,484 high quality tRNAs were identified. Blast pairwise comparison of these tRNAs against that of the domestic sheep genome (downloaded from GtRNAdb 2.0 (Chan and Lowe 2015) identified 444 tRNAs of Snow sheep that had sequence identity >95% over aligned sequence length or more than 95% of query coverage per hsp (high-scoring segment pairs) with the tRNAs of domestic sheep (Reference: Oar4.1, blastn settings: -perc_identity 0.95 -qcov_hsp_perc 0.95 –max_target_seqs 1). In addition, a total of 243 rRNAs, 1,931 snRNAs, and 782 miRNAs were also identified (Table S4).

Availability of supporting data

All the fastq files of genome sequences generated using Illumina sequencing technology and fast5 files generated using PromethION sequencing technology are uploaded on EMBL-ENA with project accession number: PRJEB38329. All the supporting files related to genome annotation can be downloaded using this link: https://doi.org/10.6084/m9.figshare.12458111.

Competing Interests

The authors declare that they have no competing interests.

Funding

This work was supported in part by the Ministry of Science and Higher Education of the Russian Federation within theme 0445-2019-0024, as well as by the Russian Foundation of Basic Research No. 18-316- 20009. This work was also supported by LMUexcellent funding of LMU Munich.

Author’s contribution
I.M., G.B., and N.Z. conceived and guided the study. G.B., I.O., and V.B. obtained the samples. S.K. and H.B. led the genome and RNA sequencing. A.D., and I.O. provided important inputs in data analysis. M.U., A.H., and E.K. analyzed the data. M.U., A.H., E.K., and A.D. wrote the manuscript. All authors read and approved the final manuscript.
References

Archibald AL, et al. 2010. The sheep genome reference sequence: a work in progress. Animal Genetics 41: 449-453. doi: 10.1111/j.1365-2052.2010.02100.x

Benson G 1999. Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Research 27: 573-580. doi: 10.1093/nar/27.2.573

Buchfink B, Xie C, Huson DH 2015. Fast and sensitive protein alignment using DIAMOND. Nature Methods 12: 59-60. doi: 10.1038/nmeth.3176

Chan PP, Lin BY, Mak AJ, Lowe TM 2019. tRNAscan-SE 2.0: Improved Detection and Functional Classification of Transfer RNA Genes. bioRxiv: 614032. doi: 10.1101/614032

Chan PP, Lowe TM 2015. GtRNAdb 2.0: an expanded database of transfer RNA genes identified in complete and draft genomes. Nucleic Acids Research 44: D184-D189. doi: 10.1093/nar/gkv1309

Chomczynski P, Sacchi N 1987. Single-step method of RNA isolation by acid guanidinium thiocyanate-phenol-chloroform extraction. Analytical biochemistry 162: 156-159. doi: 10.1006/abio.1987.9999

Deniskova TE, et al. 2018. Characteristics of the Genetic Structure of Snow Sheep (Ovis nivicola lydekkeri) of the Verkhoyansk Mountain Chain. Russian Journal of Genetics 54: 328-334. doi: 10.1134/s1022795418030031

Deniskova TE, et al. 2016. Whole genome SNP scanning of snow sheep (Ovis nivicola). Biochemistry, Biophysics, and Molecular Biology 469: 288-293. doi: 10.1134/s1607672916040141

Dohmen E, Kremer LPM, Bornberg-Bauer E, Kemena C 2016. DOGMA: domain-based transcriptome and proteome quality assessment. Bioinformatics 32: 2577-2581. doi: 10.1093/bioinformatics/btw231

Dotsev AV, et al. 2018. Genome-wide SNP analysis unveils genetic structure and phylogeographic history of snow sheep (Ovis nivicola) populations inhabiting the Verkhoyansk Mountains and Momsky Ridge (northeastern Siberia). Ecology Evolution 8: 8000-8010. doi: 10.1002/ece3.4350

Dotsev AV, et al. 2019. The first complete mitochondrial genomes of snow sheep (Ovis nivicola) and thinhorn sheep (Ovis dalli) and their phylogenetic implications for the genus Ovis. Mitochondrial DNA Part B 4: 1332-1333. doi: 10.1080/23802359.2018.1535849

Gilbert DG 2019. Genes of the pig, Sus scrofa, reconstructed with EvidentialGene. PeerJ 7: e6374-e6374. doi: 10.7717/peerj.6374

Grabherr MG, et al. 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nature biotechnology 29: 644-652. doi: 10.1038/nbt.1883
Griffiths-Jones S, et al. 2005. Rfam: annotating non-coding RNAs in complete genomes. Nucleic Acids Research 33: D121-D124. doi: 10.1093/nar/gki081

Haas BJ, et al. 2003. Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. Nucleic Acids Research 31: 5654-5666. doi: 10.1093/nar/gkg770

Haas BJ, et al. 2008. Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. Genome Biology 9: R7-R7. doi: 10.1186/gb-2008-9-1-r7

Huerta-Cepas J, et al. 2017. Fast Genome-Wide Functional Annotation through Orthology Assignment by eggNOG-Mapper. Molecular Biology and Evolution 34: 2115-2122. doi: 10.1093/molbev/msx148

Huerta-Cepas J, et al. 2018. eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. Nucleic Acids Research 47: D309-D314. doi: 10.1093/nar/gky1085

Iwata H, Gotoh O 2012. Benchmarking spliced alignment programs including Spaln2, an extended version of Spaln that incorporates additional species-specific features. Nucleic Acids Res 40: e161. doi: 10.1093/nar/gks708

Jones P, et al. 2014. InterProScan 5: genome-scale protein function classification. Bioinformatics (Oxford, England) 30: 1236-1240. doi: 10.1093/bioinformatics/btu031

Kim D, Langmead B, Salzberg SL 2015. HISAT: a fast spliced aligner with low memory requirements. Nature Methods 12: 357-360. doi: 10.1038/nmeth.3317

Korf I 2004. Gene finding in novel genomes. BMC Bioinformatics 5: 59. doi: 10.1186/1471-2105-5-59

Lagesen K, et al. 2007. RNAmmer: consistent and rapid annotation of ribosomal RNA genes. Nucleic Acids Research 35: 3100-3108. doi: 10.1093/nar/gkm160

Martin M 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal 17: 10-12.

Medvedev DG, et al. 2017. Genetic characteristics of Kodar snow sheep using SNP markers. Contemporary Problems of Ecology 10: 591-598. doi: 10.1134/s1995425517060099

Miller JM, Moore SS, Stothard P, Liao X, Coltman DW 2015. Harnessing cross-species alignment to discover SNPs and generate a draft genome sequence of a bighorn sheep (Ovis canadensis). BMC Genomics 16: 397-397. doi: 10.1186/s12864-015-1618-x

Nadler CF, Hoffmann RS, Woolf A 1973. G-band patterns as chromosomal markers, and the interpretation of chromosomal evolution in wild sheep (Ovis). Experientia 29: 117-119. doi: 10.1007/BF01913288
Nawrocki EP, Eddy SR 2013. Infernal 1.1: 100-fold faster RNA homology searches. Bioinformatics 29: 2933-2935. doi: 10.1093/bioinformatics/btt509

Parra G, Blanco E, Guigó R 2000. GeneID in Drosophila. Genome research 10: 511-515. doi: 10.1101/gr.10.4.511

Pertea M, et al. 2015. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nature biotechnology 33: 290-295. doi: 10.1038/nbt.3122

Raul Valdez JB. 1982. The Wild Sheep of the World. Mesilla: Wild Sheep and Goat International.

Rezaei HR, et al. 2010. Evolution and taxonomy of the wild species of the genus Ovis (Mammalia, Artiodactyla, Bovidae). Molecular Phylogenetics and Evolution 54: 315-326. doi: https://doi.org/10.1016/j.ympev.2009.10.037

Ruan J, Li H 2019. Fast and accurate long-read assembly with wtdbg2. 530972. doi: 10.1101/530972 %J bioRxiv

Smit A, Hubley R. 2008-2015. RepeatModeler Open-1.0, available at http://www.repeatmasker.org.

Smit AFA, Hubley R, Green P. 2013–2015. RepeatMasker Open-4.0. Seattle, Washington: Institute for Systems Biology. Available from: www.repeatmasker.org

Stanke M, et al. 2006. AUGUSTUS: ab initio prediction of alternative transcripts. Nucleic Acids Research 34: W435-W439. doi: 10.1093/nar/gkl200

Stanke M, Waack S 2003. Gene prediction with a hidden Markov model and a new intron submodel. Bioinformatics 19 Suppl 2: ii215-225. doi: 10.1093/bioinformatics/btg1080

Tarailo-Graovac M, Chen N 2004. Using RepeatMasker to identify repetitive elements in genomic sequences. Curr Protoc Bioinformatics Chapter 4: Unit 4.10. doi: 10.1002/0471250953.bi0410s05

Vaser R, Sovic I, Nagarajan N, Sikic M 2017. Fast and accurate de novo genome assembly from long uncorrected reads. Genome Research 27: 737-746. doi: 10.1101/gr.214270.116

Walker BJ, et al. 2014. Pilon: An Integrated Tool for Comprehensive Microbial Variant Detection and Genome Assembly Improvement. PLOS ONE 9: e112963. doi: 10.1371/journal.pone.0112963

Yang Y, et al. 2017. Draft genome of the Marco Polo Sheep (Ovis ammon polii). GigaScience 6. doi: 10.1093/gigascience/gix106

Zhang SV, Zhuo L, Hahn MW 2016. AGOUTI: improving genome assembly and annotation using transcriptome data. GigaScience 5: 31-31. doi: 10.1186/s13742-016-0136-3
Zhelezno-Chukotskii NK. 1994. Ecology of snow sheep of Northern Asia. In. Moscow, Russia: Nauka Press.
Figure 1. (A) Photograph of an adult *Ovis nivicola* from Suntar Hayata Ridge (Eastern Siberia). (B) Length distribution of Nanopore sequencing reads (a) and distribution of average read quality of Nanopore sequencing reads (b). (C) Visualization of the statistics of the snow sheep assembly (https://github.com/rjchallis/assembly-stats): the inner radius (highlighted in red color) represents the length of the longest scaffold, the radial axis from circumference indicates the scaffold length, the dark and light orange arcs represent the N50 and N90 scaffold lengths respectively. The outermost circular layer (in shades of blue) shows the base composition at the given coverage (represented in terms of percentage) of the genome.
Figure 2. (A) Interspersed repeat landscape of snow sheep genome. (B) Coverage of *Ovis aries* reference (*Oar_rambouillet_v1.0*) by snow sheep genome assembly. The X-axis represents the position and Y-axis represents chromosome number of *Ovis aries* reference. The red bars represents the size of the chromosomes of Ovis aries reference.