High quality genome of the Mountain hare (*Lepus timidus*) with new chromosome level scaffolding

Craig Michell[1,2], Joanna Collins[3], Zsófia Fekete[4], Riikka Tapanainen[1], Jonathan M. D. Wood[3], Steffi Goffart[1] and Jaakko L. O. Pohjoismäki*[1]

[1]University of Eastern Finland, Department of Environmental and Biological Sciences – Joensuu, Finland
[2]King Abdullah University of Science and Technology, Red Sea Research Center, Biological and Environmental Science and Engineering Division – Thuwal, Kingdom of Saudi Arabia
[3]Tree of Life, Wellcome Sanger Institute, Hinxton, Cambridge, UK.
[4]Hungarian University of Agriculture and Life Sciences, Institute of Genetics and Biotechnology, Gödöllő, Hungary

*Corresponding author
Correspondence: jaakko.pohjoismaki@uef.fi

Abstract

Here we present a high-quality genome assembly of the mountain hare (*Lepus timidus* L.), based on a fibroblast cell line of a male specimen from Ilomantsi, Eastern Finland. The mountain hare represents the first Finnish contribution to the European Reference Genome Atlas pilot effort to generate reference genomes for European biodiversity.

The genome was assembled using 25X PacBio HiFi sequencing data and scaffolded utilizing a Hi-C chromosome structure capture approach. After manual curation, the assembled genome length was 2,930,972,003 bp with N50 125.8Mb. 93.16% of the assembly could be assigned to 25 identified chromosomes (23 autosomes plus X and Y), matching the published karyotype. The chromosomes were named according to size. The genome has a high degree of completeness based on the BUSCO score (mammalia_odb10 database), Complete: 96.1% [Single copy: 93.1%, Duplicated: 3.0%], Fragmented 0.8%, and Missing 2.9%. The mitochondrial genome of the cell line was sequenced and assembled separately.

Compared to the previous pseudoreference genome of *L. timidus* ssp. *hibernicus* Bell, assembled using the rabbit genome, this new reference genome represents the nominal subspecies and the species-specific chromosomal conformation. In addition to providing a useful comparison for population studies, the genome can offer insight into the chromosomal evolution

among Glires in general and Lagomorpha in particular. The chromosomal assembly of the genome also demonstrates that the cell line has not gone through karyotypic changes during culture.

Keywords: Lagomorpha, reference genome, chromosome structure, phylogeny, fibroblast cell line

Introduction

The mountain hare (*Lepus timidus* L.) is the most widespread hare species in the world, with its distribution ranging from Ireland to Japan, including several subspecies (Angerbjörn & Schai-Braun, 2022). The species is adapted to the cold and snowy conditions of northern Eurasia by having wide snowshoe feet and white winter pelage. The current climate change-induced shortening of the snow-covered season is detrimental for the mountain hare, especially due to camouflage mismatch (Zimova *et al*, 2020; Zimova *et al*, 2016), resulting in range contraction to the north or higher altitudes in mountains, while benefitting the expansion of its more southern relative, the brown hare (*L. europaeus*) (Jansson & Pehrson, 2007; Levanen *et al*, 2018a; Rehnus *et al*, 2018; Reid, 2011; Thulin, 2003).

The mountain hare has shown resilience during the past warm periods of the Holocene (Smith *et al*, 2017), diverging also to more temperate climate-adapted subspecies (Giska *et al*, 2022), such as the Irish hare (*L. t. hibernicus* Bell) and the heath hare (*L. t. sylvaticus* Nilsson), which quite likely represent relatively old evolutionary lineages, genetically distinguishable from the widespread nominal subspecies (Giska *et al.*, 2022; Michell *et al*, 2022). In addition, post-glacial contact with other congeneric hare species has resulted in recurrent hybridization and genetic introgression (Levanen *et al*, 2018b; Melo-Ferreira *et al*, 2009), some of which has apparently adaptive significance (Giska *et al*, 2019; Giska *et al.*, 2022; Pohjoismaki *et al*, 2021). Besides the fact that especially the geographically restricted, relic-like subspecies are highly threatened (Caravaggi *et al*, 2017; Thulin *et al*, 2021) and thus interesting in the context of conservation genetics, hares provide ample opportunities to study local adaptations (Giska *et al.*, 2022; Michell *et al.*, 2022), genomic introgression (Giska *et al.*, 2019; Levanen *et al.*, 2018b; Pohjoismaki *et al.*, 2021) and genetic constituents of species boundaries (Gaertner *et al*, 2022). These studies benefit greatly from the increased availability of genomic information.

Over the last decade there has been a boom in sequencing technologies. This has led to many sequencing initiatives, such as the European Reference Genome Atlas and he Darwin Tree of Life, that aim to produce high quality reference genomes for a variety of organisms (Blaxter *et al*, 2022). Once such technological advance is third generation sequencing or high-throughput sequencing of long molecules, providing the core of many of these assemblies. With read length N50's > 19 kbp, many genomes are being released with high contiguity compared to the early 2000's that were based on second generation short read sequencing technologies. Similarly, Hi-C techniques have further propelled these genomes into the realm of chromosome-scale assemblies, a level previously only held by model organisms. Coupled with these advances is also a decrease in price per base pair, making sequencing available to many laboratories and research groups. Combining these two technologies, we can assemble reference genomes for non-model organisms (Lawniczak *et al*, 2022). Here we apply these two technologies to produce a high-quality reference genome of the nominal mountain hare subspecies (*Lepus timidus timidus*).

While there is an existing genome assembly for the mountain hare (GCA_009760805), it represents the Irish subspecies *L. t. hibernicus* and is based on a female specimen, thus lacking the Y-chromosome. Furthermore, the existing assembly is a so-called "pseudoreference", scaffolded using the rabbit (*Oryctolagus cuniculus* L.) reference genome (Marques *et al*, 2020). The original rabbit reference genome had been established using whole-genome shotgun sequencing of females representing the inbred Thorbecke New Zealand White line and was quite fragmented (Carneiro *et al*, 2014), updated only recently with long-read sequencing data (Bai *et al*, 2021). In addition, the diploid chromosome number in domestic rabbits is 2n = 44 (Korstanje *et al*, 1999), whereas it is 48 in the mountain hare (Gustavsson, 1972).

To obtain a reference genome conforming to current standards, we used a fibroblast cell line of a male specimen from Ilomantsi, Eastern Finland (Gaertner *et al.*, 2022), as a source of high-molecular weight (HMW) DNA and fresh RNA. Compared to solid tissues, fibroblasts are optimal for Hi-C, as the method was originally developed for cells growing in a monolayer (Lieberman-Aiden *et al*, 2009). The specimen was collected outside of the brown hare range (Levanen *et al.*, 2018a; Levanen *et al.*, 2018b) to avoid the influence of hybridization. Furthermore, the sampled population represents a continuum across the vast eastern Taiga zone, making our sample very representative geographically for the nominal mountain hare subspecies. It is worth noting that Linné did not formally designate a type specimen for the mountain hare, and the type locality has been later decided to be Uppsala in Sweden (Angerbjörn & Schai-Braun, 2022). The Swedish hares belong to the same larger Fennoscandian population as our specimen, which is important, as sampling from or near the type locality is considered desirable for the representativeness of the reference genome (Lawniczak *et al.*, 2022).

We assembled the genome of *L. timidus* using 25X genome coverage of PacBio HiFi read data and then further scaffolded the assembly with Hi-C sequencing data. The assembled genome is 2.96 Gbp in length with a contig N50 of 58 Mbp and a scaffold N90 of 25. The genome has a high degree of completeness based on the BUSCO score, Complete: 96.1% [Single copy: 93.1%, Duplicated: 3.0%] Fragmented 0.8%, and Missing 2.9% based on the mammalia_odb10 database. As the scaffold N90 is 25, the genome assembly is near-chromosome scale as there is an expected 23+X+Y karyotype for this species. By using RNA-seq data as extrinsic evidence for the genome annotation, we were able to produce 19,906 gene models for this assembly. We believe that this genome will provide a good companion to the currently available *L. timidus* genome, as it has a different evolutionary background and can be included together with the current assembly in future phylogenomic studies of hares.

Methods

Sampling and confirming of the species identity

A young male mountain hare was hunted in September 2019 in Kelovaara, Ilomantsi, Finland (63.0609 N, 30.6248 E). The sampling had minimal impact on the local mountain hare population and no impact on the habitat. As mountain hares are legal game animals in Finland and the hunting followed the regional hunting seasons and legislation (Metsästyslaki [Hunting law] 1993/615/5§), the sampling adheres to the ARRIVE guidelines and no ethical assessment for the sampling was required. Sampling also did not involve International Trade in Endangered

Species of Wild Fauna and Flora (CITES) or other export of specimens, as defined by the Convention on Biological Diversity (CBD).

The species identity was confirmed at site based on the morphological features distinguishing the mountain hare from the brown hare, the only other hare species in northern Europe. The collection location is remote wilderness, dominated by boreal Taiga Forest (Figure 1A) and devoid of brown hares, which in Finland are dependent on man-made environments (Levanen *et al.*, 2018a). Further analysis of the coding part of the genome and mitochondrial DNA haplotyping showed no ancestral admixture with brown hares (Gaertner *et al.*, 2022).

Generation and vouchering of the cell line

The fibroblast cell line (LT1) was isolated from the specimen as described earlier (Gaertner *et al.*, 2022) and is deposited under voucher number ZFMK-TIS-69753 into the biobank of the Stiftung Leibniz-Institut zur Analyse des Biodiversitätswandels, Zoological Research Museum Alexander König (ZFMK), Bonn, Germany.

HMW DNA extraction and PacBio HiFi sequencing

A salting out-method was used to extract high molecular weight DNA from cells grown to confluency on a 10 cm cell culture dish. Briefly, the cells were detached from the dish using trypsin, followed by centrifugation and two washes with PBS. The salting out-method then followed the 10X Genomics “Salting Out Method for DNA extraction from cells” protocol. The high molecular weight DNA was quantified using the Qubit 3.0, followed by qualification with an 0.8 % Agarose gel. The DNA was analyzed by DNA Sequencing and Genomics Laboratory, Institute of Biotechnology, University of Helsinki on the PacBio Sequel II. Sequence data from two flow cells was produced.

Mitochondrial DNA (mtDNA) sequencing

The mitochondrial genomes of our hare cell lines have been sequenced for the purposes of another study by PCR-amplifying the mtDNA in approximately 2 kb overlapping fragments. The primers used to amplify the mountain hare mtDNA were as follows:

Lt93F: TTGTTTCGTAGCAAGTTTACACATGC
Lt184R: GCTTAATACCTGCTCCTCTTGATCAG
Lt1580F: TTAAACCCATAGTTGGCCTAAAAGC
Lt1635R: TTGAGCTTTAACGCTTTCTTAATTGA
Lt3045F: AGGCGTATTGTTTATCTTAGCAACCT
Lt3175R: CTTCGTATGAAATGGTTTGTGCAA
Lt3921F: CCCCCTAATCTTTTCCATCATTCTAT
Lt4482R: TCATCCTATATGGGCAATTGAGGAAT
Lt4689F: AGGCTTTATCCCAAAATGAATCATCATTCA
Lt5417R: AGGCTCCAAATAAAAGGTAGAGAGTT
Lt6696F: ACACCGTTTCATCGATAGGCTCATTC
Lt6756R: ATAAAGATTATTACTATTACAGCGGTTAGA
Lt8603F: AGCCTATATTTACACGATAACACCTAATGA
Lt8698R: CGGATAGGGCTCCAGTGAGTGG

Lt10552F: TTGAAGCAACACTAATTCCTACACTA
Lt10613R: TCGTTCTGTTTGGTTACCTCATCGT
Lt11301F: ACCATTAATCTTCTAGGAGAGCTTCT
Lt11807R: AGGATAATGATTGAGACGGCTATAGA
Lt12407F: GTCTTATCCTAGCCGCTACAGGTAAA
Lt12791R: GAGCATAAAAAGAGTATAGCTTTGAA
Lt14204F: ATTGTTAACCACTCCCTAATTGACCT
Lt14514R: CCAATATTTCAGGTTTCTAGGTAAGT
Lt16056F: TGGGGTATGCTTGGACTCAAC
Lt16119R: TCGTCTACAATAAGTGCACCGG

In total, 12 separate reactions were prepared to cover the mitochondria genome:

1. Lt16056F + Lt184R: 1852 bp*
2. Lt93F + Lt1635R: 1543 bp
3. Lt1580F + Lt3175R: 1596 bp
4. Lt3045F + Lt4482R: 1438 bp
5. Lt3921F + Lt5417R: 1497 bp
6. Lt4689F + Lt6756R: 2068 bp
7. Lt6696F + Lt8698R: 2003 bp
8. Lt8603F + Lt10613R: 2011 bp
9. Lt10552F + Lt11807R: 1256 bp
10. Lt11301F + Lt12791R: 1491 bp
11. Lt12407F + Lt14514R: 2108 bp
12. Lt14204F + Lt16119R: 1916 bp

(*Expected fragment size based on the published *Lepus timidus* mtDNA sequence from Finland [NC_024040] (Melo-Ferreira *et al*, 2014). The true fragment size is dependent on the haplotype. See the discussion).

The fragments were amplified from total DNA preparations using a PCR program with 1 min 94°C denaturing step, followed by 35 cycles of 94°C for 15 s, 56°C for 15 s and 72°C for 2 min, as well as by a final 3 min elongation step at 72°C. The obtained products were gel purified using the GeneJET gel extraction kit (Thermo Scientific) and sent for sequencing using Illumina MiniSeq at the Genome Center of Eastern Finland.

The sequence of the non-coding region containing PCR fragment (Lt16056F + Lt184R) was further validated by Sanger sequencing, applying also the following additional primers:
Lt680F: AGAACCGTGACATAGCACTTACTTTC
Lt945R: TAACATATTGGTGTTAGAATGTTTTTAGT

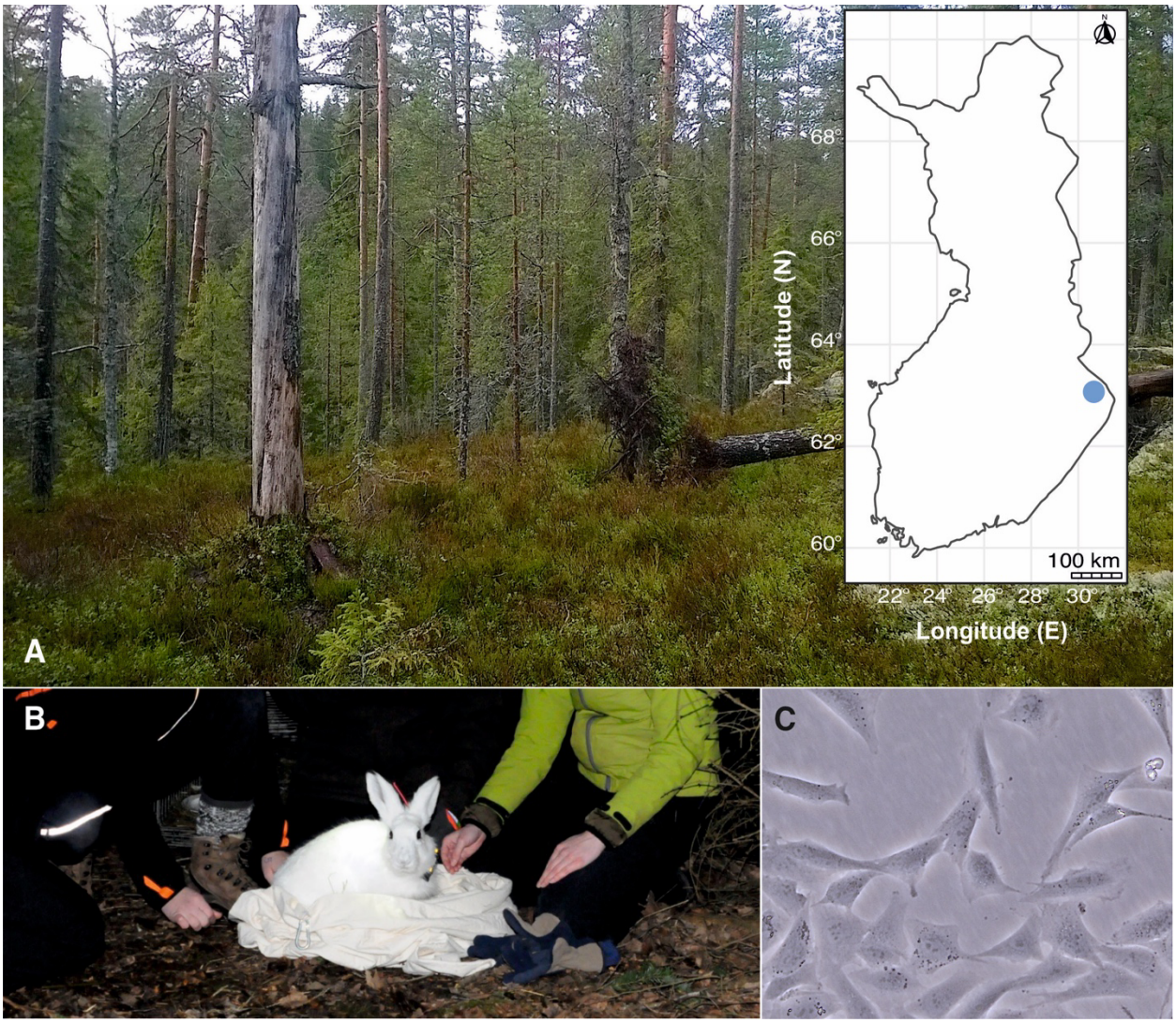

Figure 1. A) The geographic location and a view of the sampling biotope in Ilomantsi. Photo by Jaakko Pohjoismäki. B) Although the cell line used in this study originates from a killed specimen, fibroblasts can be isolated from e.g. ear clippings taken during tagging of animals for behavioral studies. Photo by Mervi Kunnasranta. C) The genome presented here was generated using the LT1 cell line, exhibiting a typical fibroblast morphology.

Hi-C library preparation

Hi-C sequencing libraries were prepared following the protocol of (Belaghzal *et al*, 2017) with the following changes: 1.) To prepare a diverse sequencing library, we performed the Hi-C protocol in triplicates, 2.) Size fractionation was performed using NEBNext Ultra II FS DNA module, 3.) Library preparation was performed using the NEBNext Ultra II library preparation kit for Illumina and 4.) Library enrichment was performed using triplicate PCR reactions with six cycles of PCR. The PCR reactions were then purified using Ampure XP beads at a ratio of 0.9X. The final clean libraries were quantified using the Qubit, followed by agarose gel electrophoresis to confirm the fragment size. The sequencing was performed on a single lane of the Illumina NovaSeq 6000 using the SP flowcell with paired end chemistry 2 x 150bp.

Genome assembly

HiFiasm version 0.18.7 (Cheng *et al*, 2021) was used to assemble the 25X (rq >0.99) PacBio HiFi reads using the arguments *-l3 –h1 and –h2 –primary* to integrate the Hi-C read data and produce a primary assembly. We then continued with the scaffolding of the primary assembly and annotation. To process the HiC data, we first mapped the HiC data to the primary genome assembly using BWA-mem version 0.7.17 (Li & Durbin, 2009) with the arguments *-SP*. The mapped reads were then parsed and filtered using pairtools version 1.0.2 (Open2C *et al*, 2023). To parse the aligned Hi-C reads, we used the options *–no-flip –dd-columns mapq –walks-policy mask*. The parsed pair data was then sorted and deduplicated using default arguments. Finally, we selected unique-unique and unique-rescued pairs and split these into the pairs file and bam file for input in YaHS version 1.1 (Zhou *et al*, 2023). YAHS was run using the default parameters with the primary contig assembly and the filtered Hi-C bam file. Contiguity and general genome statistics were calculated using QUAST version 5.2.0 (Mikheenko *et al*, 2018). We assessed the completeness of the genome by calculating the number of complete single copy orthologs with BUSCO version 5.1.2 (Manni *et al*, 2021), using the mammalia_odb10 database as well as the more lineage specific glires_odb10 database.

Genome annotation

Repeat annotation of the genome was performed with EDTA version 2.1.0 (Ou *et al*, 2019), a *de novo* repeat identification pipeline. Using the repeat library produced by EDTA, we masked the scaffolded genome using RepeatMasker version 4.1.1 (Smit *et al*, 2013-2015). RNA-seq data from the same cell line (SRA accession number: SRR18740842) as well as RNA-seq data from other *L. timidus* libraries (SRA accession number: SRR10020054, SRR10020055, SRR10020060, SRR10491719, SRR18740839, SRR18740840, and SRR18740841) was collected from the sequence read archive (SRA). The RNA-seq data was trimmed using fastp version 0.23.2 (Chen *et al*, 2018) and mapped against the masked genome using HISAT2 version 2.1.0 (Kim *et al*, 2019) with the default parameters. Furthermore, we included the protein sequences of the assembled transcripts from this cell line as further evidence for the genome annotation. These lines of gene evidence were included in the annotation using BRAKER3 version 3.02 (Bruna *et al*, 2021). Telomeric sequences [AACCCT]n were identified using A Telomere Identification toolkit (tidk) version 0.2.31. The telomeric sequence copy number was then calculated in windows of 200kb for visualization using Circos (Krzywinski *et al*, 2009).

Mitochondrial DNA assembly and annotation

Mitochondrial DNA was assembled from the PCR-amplified and Illumina sequenced mtDNA using the MitoZ pipeline (Meng *et al*, 2019). After comparing the results of the pipeline's outputs with different kmer options, we selected the best assembly as final. Run options used in the final assembly were --clade Chordata –fastq_read_length 150, --requiring_taxa Chordata --genetic_code 2 --kmers_megahit 21 29 39 59 79 99 119 141. The tools invoked by the pipeline included fastp (Chen *et al.*, 2018) for cleaning the raw data, MEGAHIT (Li *et al*, 2015) for assembly, after which sequences were filtered to ensure the correct taxa by HMMER (Wheeler & Eddy, 2013) and further filtered for completeness of protein-coding genes. Annotation steps were done using TBLASTN (Gertz *et al*, 2006), GeneWise (Birney *et al*, 2004) and MiTFi (Juhling *et al*, 2012). Final

manual curation, the annontation of the non-coding region (NCR) as well as the illustration of the mitochondrial genome was done using Geneious® 10.2.6. The functional loci on the NCR were identified based on the similarity with the human (NC_012920) and mouse (FJ374652) NCR sequences.

Manual curation

The assembled and annotated genome was manually curated to further improve its quality as described in (Howe *et al*, 2021).

Comparison with previous assembly

We performed a comparison of our scaffolded assembly to the current *L. timidus* genome assembly (NCBI Accession number: GCA_009760805). Mapping to the genome was performed using minimap2 version 2.21 (Li, 2018) with the arguments *-asm5*. A dot plot of the alignment was created using the R script pafCoordsDotPlotly.R (https://github.com/tpoorten/dotPlotly).

Results

The genome assemblies can be accessed via BioSample accession SAMN33984520 as well as BioProject accession PRJNA1009711 for the primary assembly and PRJNA1009710 for the alternative assembly.

Genome assembly

The expected haploid genome size of the *L. timidus* is 3.25 pg (3,178,500,000 bp) (Vinogradov, 1998), containing 23 autosomal chromosomes and two sex chromosomes (Gustavsson, 1972). PacBio HiFi sequencing with two flow cells resulted in 25 X coverage of the expected genome size of *L. timidus*. The sequence N50 of the HiFi data was 19.97 Kb. Based on the PacBio HiFi data, the expected genome size using k-mer (K = 21) coverage is 2.4 Gbp (Figure 2A). The illumina sequencing of the Hi-C data produced 494,285,585 paired reads representing about 51 X coverage of the genome. The duplication rate of the Hi-C data was 17 %. Assembly with HiFiasm yielded a contig assembly of 2.96 Gbp made up of 671 contigs with a contig N50 of 58 Mbp. The longest contig was 149 Mbp. Using the uniquely mapped Hi-C data, we were able to scaffold the contigs and fix misassembled contigs. The Hi-C scaffolded assembly (Figure 2B) was 2.96 Gbp in size, similar to the contig assembly, the largest scaffold was 149 Mbp. The scaffold N50 was 124 Mbp (N=10) and the scaffold N90 was 21 Mbp (N=25). The BUSCO scores of the *L. timidus* assembly suggest that it is near-complete, with the following results: Complete: 96.1% [Single copy: 93.1 %, Duplicated: 3.0 %] Fragmented 0.8 %, and Missing 2.9 % based on the mammalia_odb10 database.

To further improve the assembly, manual curation was performed using Hi-C maps (Table 1) and resulted in 76 scaffold breaks, 81 joins and removal of 62 erroneously duplicated contigs. These interventions led to an increase in scaffold N50 of 1.69 % from 123.7 Mb to 125.8 Mb, and a reduction in scaffold count of 9.21 % from 716 to 650. Of the finalised assembly, 93.16 % could be assigned to 25 identified chromosomes (23 autosomes plus X and Y) matching the published

karyotype. Chromosomes were named according to size. The curated genome length was 2,930,972,003 bp.

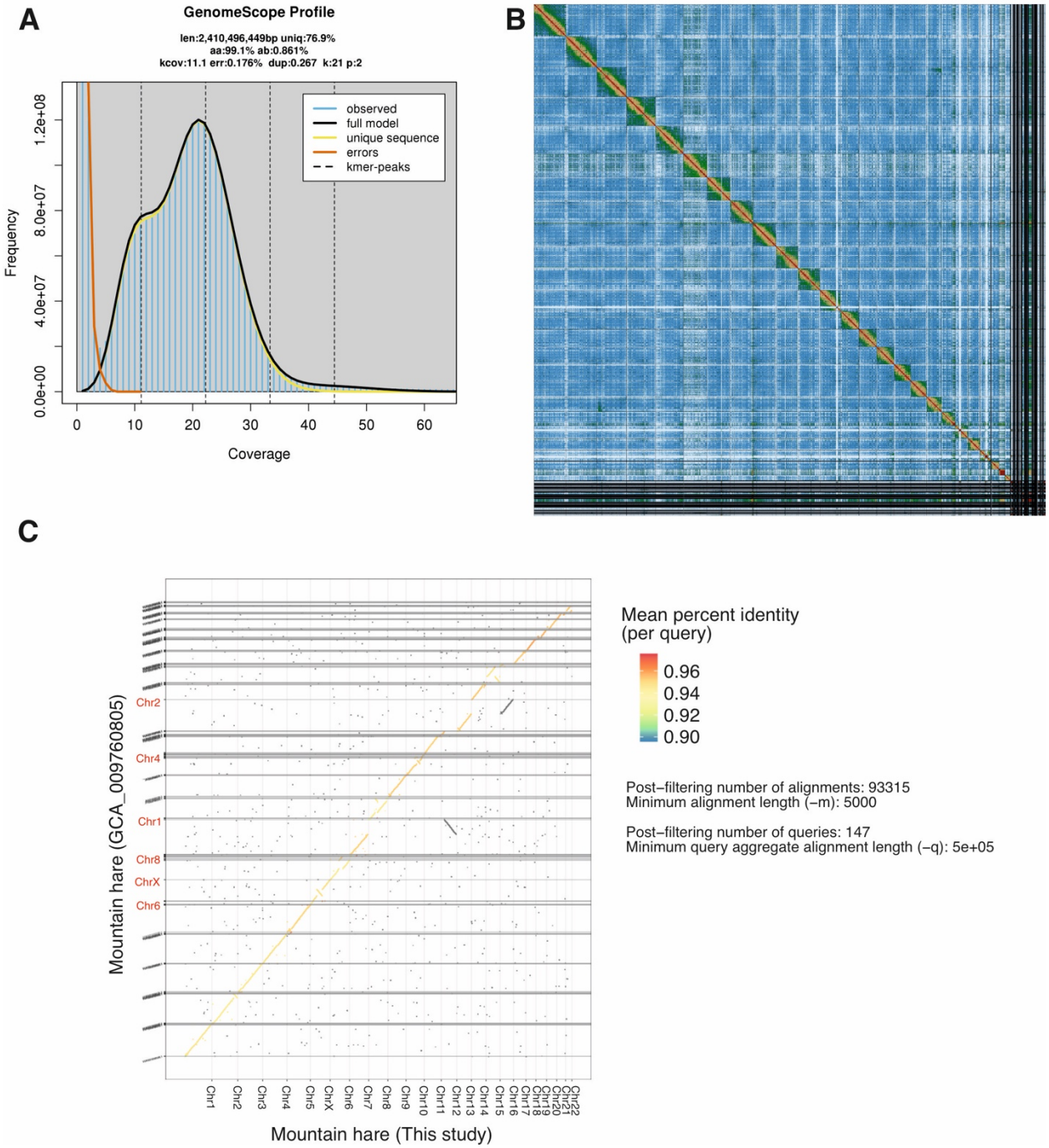

Figure 2. A) Genomescope2 profile of the PacBio HiFi data. B) Pretext map of Hi-C scaffolding after manual curation. C) Dot plot comparison with the previous assembly. Some examples of the previous chromosome assemblies are highlighted.

319 **Table 1.** Mountain hare reference genome chromosome assignment and assembly statistics

Chromosome assignment	
Chr length	2,730,543,680
Chr length %	93.16
Autosomes	23
X	1
Y	1
Unlocalised	19

Statistics for primary assembly after manual curation	
# scaffolds	650
Total scaffold length	2,930,972,003
Average scaffold length	4,509,187.70
Scaffold N50	125,776,599
Scaffold auN	117,151,857.99
Scaffold L50	10
Largest scaffold	184,151,585
Smallest scaffold	1000
# contigs	1068
Total contig length	2,930,889,045
Average contig length	2,744,278.13
Contig N50	43,409,466
Contig auN	47,572,960.95
Contig L50	21
Largest contig	141,329,000
Smallest contig	1000
# gaps in scaffolds	418
Total gap length in scaffolds	82,958
Average gap length in scaffolds	198.46
Gap N50 in scaffolds	200
Gap auN in scaffolds	199.70
Gap L50 in scaffolds	208
Largest gap in scaffolds	200
Smallest gap in scaffolds	31
Base composition	
A	821,587,687
C	645,897,412
G	644,604,632

T	818,799,314
GC content	44.03 %
# soft-masked bases	0
# segments	1,068
Total segment length	2,930,889,045
Average segment length	2,744,278.13
# gaps	418
# paths	650

Genome annotation

EDTA produced a curated custom repeat library of 7,045 repetitive elements. Interestingly, the proportion of the genome masked as repetitive elements was higher than expected at 46.8 % of the genome. In the previous genome assembly, only 28 % of the genome was masked, and similarly the *k*-mer based estimate of the repetitive elements was 26.7 %. To annotate the genome, we mapped ~1.6 billion RNA-seq reads from previously published RNA-seq libraries, including this cell line. The average mapping rate of all libraries was 89.95 %, which we consider a good indicator of the quality of this genome. Using BRAKER3, we were able to annotate 19,906 gene models in the genome. We then ran BUSCO on the predicted genes and the results show a good level of completeness Complete: 81.1 % [Single copy: 63.3 %, Duplicate: 17.8 %], Fragmented: 1.1 %, Missing: 17.8 %, Total: 9226 when compared to the mammalia_odb10 database. Telomeric sequences on both ends of the chromosome were detected on 13 chromosomes (Figure 3A). Furthermore, telomeric sequences were found in high copy number throughout the chromosome length, a feature previously noted using FISH in mountain hares (Forsyth *et al*, 2005).

Comparison to previous assembly

Minimap2 was able to align 97.3 % of the contig sequences from the previous genome assembly to this genome (Figure 2C). The previous genome assembly was scaffolded against the rabbit reference genome, and it represents an evolutionary distinct lineage to the Finnish mountain hare sequenced in this study (Giska *et al.*, 2022). Despite these differences, we have a high level of sequence similarity with the previous genome assembly as well as a high degree of synteny. However, there a few minor notable inversions and structural differences between the previous genome and the one presented here (Figure 2C). These differences are mostly explained by the fact that the previous genome assembly retains the chromosomal arrangement of the rabbit genome (Beklemisheva *et al*, 2011). In assembly presented in this study, chr7 and chr 12 have a good degree of synteny with the Chr1 (Chr1 - Rabbit) of the previous assembly, demonstrating that the rabbit chromosome 1 is a fusion of two ancestral lagomorph chromosomes. Similarly, also the previous assembly Chr2 – Rabbit is a fused chromosome and has good synteny and orientation with chr13 and chr16 in this genome. This exemplifies the importance of high quality reference genomes to analyze chromosome evolution across lineages and opens up doors for future studies of chromosomal evolution among mammals.

Mitochondrial genome

The assembly of the Illumina data provided a 16,922 bp mtDNA sequence, which was complete and circular, with a 142 bp non-repetitive overlap between the 3´ and 5´ ends of the original assembly, which was then cut from the final version. All the expected vertebrate mitochondrial genes, 37 in total, were found (Figure 3B).

Interestingly, the primary assembly was notably smaller than was expected based on the original PCR amplicons, with a notable difference between the expected and obtained non-coding region (NCR) length. Sanger sequencing of the PCR-fragment containing the NCR (Lt16056F + Lt184R) revealed that it contained a sequence run consisting of five head-to-tail repeats of a 140 bp element between the OriH and tRNA-Phe, explaining the incorrect assembly of the NCR sequence from the short-read Illumina data. With the addition of the repeat elements, the final mitochondrial genome size of our specimen is 17,482 bp.

Based on the sequence similarity with human and mouse mitochondrial genomes, we were able to tentatively identify the termination associated sequence (TAS, 2 nt difference to human), required for the generation of the 7S DNA and thus the displacement loop (D-loop) structure on the NCR (Crews *et al*, 1979). Interestingly, if assuming that the 7S DNA is about the same size as in other mammals (600-650 nt) its 5´-end would map closely at a previously identified tRNA-like cloverleaf-structure on the mountain hare mtDNA NCR (Melo-Ferreira *et al.*, 2014). We have preliminary assigned this structure to correspond to the heavy-strand origin of replication (OriH), supported by the fact that a similar, tRNA sequence-derived hairpin structure is required for the priming of the origin of the light-strand replication (OriL) (Fuste *et al*, 2010). Experimental validation of the proposed OriH is required.

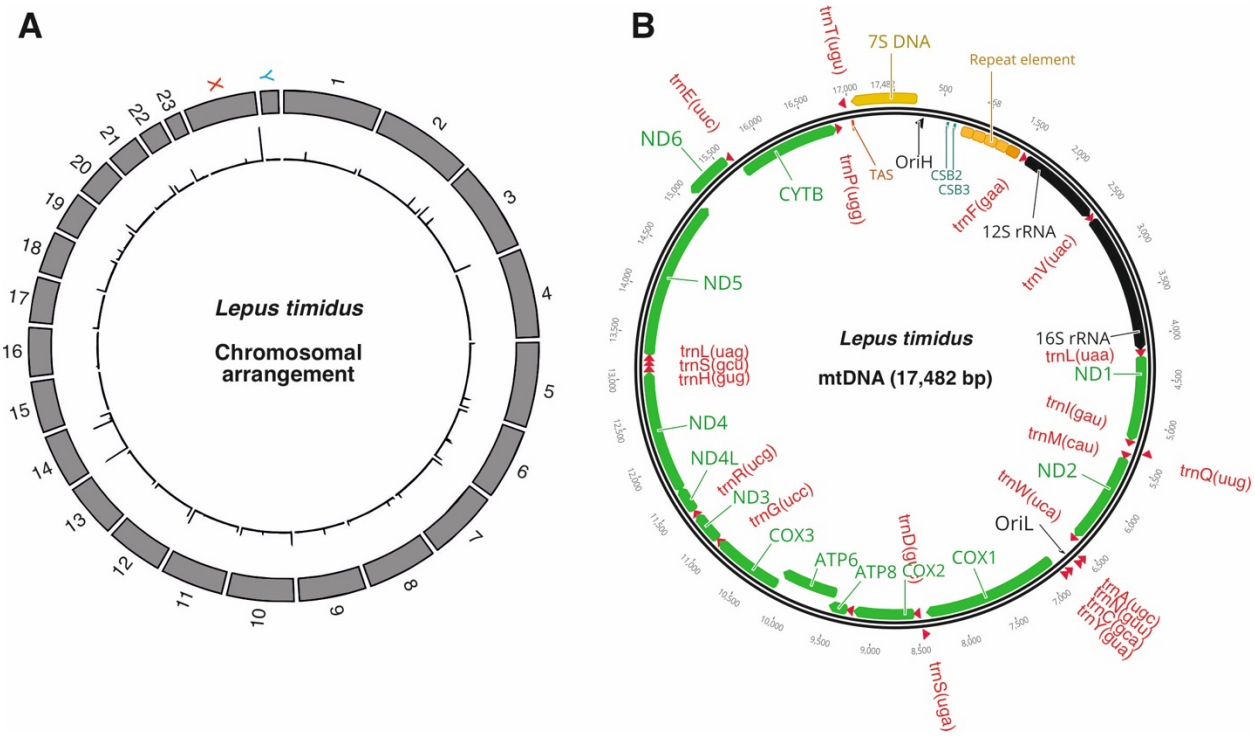

Figure 3. A reference genome for the mountain hare. A) A size map of 23 autosomes as well as X and Y sex chromosomes of *Lepus timidus*. Telomeric sequences and their relative lengths are

381 indicated on the inner circle. B) A schematic illustration of the mountain hare mitochondrial
382 genome. Genes encoded by the H-strand are placed inside and the L-strand genes outside of the
383 circle. NCR = non-coding region.

384 Discussion

385

386 Given the challenges of assembling a large mammalian genome, we have produced a high-quality
387 genome of the mountain hare. The genome has a high degree of contiguity (Contig N50 43 Mbp)
388 and completeness (BUSCO complete 96.1 %). Interestingly, the genome contains a high number
389 of repetitive elements, a fact that warrants further investigation to elucidate their identity. The
390 genome is in chromosome scale, and all 23 autosomes as well as X and Y chromosomes could be
391 scaffolded. Furthermore, the genome has a high degree of similarity to the previously published
392 genome but is of higher quality. It is not only much more contiguous but also resolves
393 chromosome structure including telomeric repeats on both ends of several chromosomes (Figure
394 3A).

395 Together, this genome and the previously published *L. timidus* genome will provide a solid
396 base for future studies. This has important implications for the fields of evolutionary and
397 conservation biology. Due to the increase in contiguity, it will be possible to perform accurate
398 analyses of linkage disequilibrium as well as the identification of runs of homozygosity (ROH).
399 Accurate ROH analysis is useful for identifying and understanding the impact of inbreeding burden
400 on the genome of increasingly threatened and isolated species, as quantitative traits can be
401 associated with ROH burden (Ceballos *et al*, 2018). Furthermore, this information might provide
402 important insights into the conservation of other threatened hares, such as the heath hare (*Lepus*
403 *timidus sylvaticus*) and Irish mountain hare (*Lepus timidus hibernicus*).

404 The obtained 17,482 bp long mitochondrial genome also provides some interesting
405 aspects to the mtDNA sequence variation within mammalian species. The size difference to
406 human (16,569 bp) and mouse (16,298 bp) mitochondrial genomes is caused by a rather long
407 non-coding region (2,034 bp vs. 1,123 bp and 879 bp in humans and mice respectively), longer
408 rRNA and tRNA genes, as well as additional non-coding nucleotides between genes. Interestingly,
409 the mitochondrial genome of our specimen is slightly smaller than the 17,755 bp long previous
410 NCBI Reference Sequence for mountain hare (NC_024040.1), also from Finland (Melo-Ferreira *et*
411 *al.*, 2014). The main size difference can be attributed to copy number difference in the repeat
412 elements on the non-coding region with seven present in the previous reference mtDNA vs five
413 in our specimen. Indels elsewhere in the NCR explain why the length difference does not add up
414 to simple multiplies of 140 bp repeat element. We have noted similar within species size variation
415 in the NCR among Finnish mountain hares as published earlier (Melo-Ferreira *et al.*, 2014), with
416 some evidence of these variants occurring in multiple heteroplasmy within an individual (a
417 manuscript in preparation). To our understanding, such within population and within individual
418 length variation in the non-coding region is unusual and warrants for further investigation.

419 The main technical challenge for highly continuous, good quality reference genomes is the
420 availability of relatively large quantities of intact HMW DNA for HiFi sequencing. This typically
421 requires snap freezing of tissue samples and assuring an intact cold chain of below -80 °C for their
422 sending, handling, and long-term storage. While development in sequencing technologies can
423 reduce the required DNA amounts, obtaining high sequencing coverage over a large vertebrate

genome is easier when material is plentiful. Similarly, RNA is very sensitive to degradation in post-mortem tissue samples, also requiring immediate preparation or snap-freezing to be suitable for RNA-seq. Although transcript data is not obligatory, it is highly useful for the purpose of genome annotation. The requirement of large chunks of fresh tissue for -omics purposes complicates the sampling of most vertebrates species. For example, in Finland all land vertebrates are protected by law, with very few exceptions, such as game animals and pest species. Capturing and lethal sampling is especially problematic for endangered species. Not only will the sampling require extensive permits but is also difficult to justify and conduct ethically.

Although the source specimen for the cell line utilized in this study was hunted, fibroblasts can be isolated from relatively small skin biopsies, such as ear clippings (Seluanov *et al*, 2010), with minimally invasive sampling and harm to the individual (Figure 1B). These cells can be expanded in tissue culture and, when immortalized, provide a highly scalable source of fresh DNA and RNA. Furthermore, living cells can be stored as frozen stocks in biobanks for decades (Fazekas *et al*, 2017). Tissue sampling can be done directly into the growth medium in ambient temperature, allowing several days for the transport to the laboratory without a need for dry ice or N_2-cryocontainers, which greatly simplifies collection and logistics. The same applies to the isolated cells. As example, we sent living cell suspensions from Joensuu, Finland to the ZFMK biobank in Bonn, Germany, over regular airmail in ambient temperature. It is worth to point out that the obtained genome scaffolded to 23 autosomes, X and Y chromosomes, as expected for the species, showing no evidence of karyotypic drift caused by cell culture, dispelling possible concerns for the use of cell lines for such work (Gaertner *et al.*, 2022).

While the mountain hare is an important game species throughout most of its range, it is also an iconic mammalian species of the boreal Eurasian Taiga zone and an important part of the local food webs. Consequently, a lot of work has been already done in terms of population genetics, evolution and genomics of mountain hares [e.g. (Ferreira *et al*, 2021; Gaertner *et al.*, 2022; Giska *et al.*, 2022; Levanen *et al.*, 2018b; Marques *et al.*, 2020; Melo-Ferreira *et al.*, 2009; Michell *et al.*, 2022; Pohjoismaki *et al.*, 2021; Smith *et al.*, 2017)]. However, a high-quality reference genome will allow us to peer deeper into genome structure and chromosome evolution, which will be particularly interesting in the context of genetic introgression (Juric *et al*, 2016) and maintenance of species boundaries. Any issues of genetic compatibility should leave their fingerprint in the genomes (Skov *et al*, 2023), providing interesting insight into the speciation mechanisms (Wolf *et al*, 2010). Other *Lepus* reference genomes available soon will help to complete this picture.

Acknowledgements

We acknowledge the DNA Sequencing and Genomics Laboratory, Institute of Biotechnology, University of Helsinki for the PacBio and Hi-C sequencing. Dr Seppo Helisalmi and Dr Joose Raivo from the Gene diagnostics lab at Genome Center of Eastern Finland are thanked for providing the mtDNA sequencing service. We are grateful to the European Reference Genome Atlas (ERGA) for the opportunity to include our mountain hare genome into their pilot project. Dr Ann McCartney is thanked for her kind help throughout the process and Mr Jussi Mähönen and Mr Manu Soininmäki for their contribution in obtaining the mountain hare specimen. We thank Ms Anita Kervinen for her valuable laboratory assistance and Dr. Mervi Kunnasranta for kindly allowing to use her mountain hare photo in the publication.

Data, scripts, code, and supplementary information availability

The genome assembly is available from the NCBI database under BioProject ID PRJNA950335. The cell line is available through the corresponding author as well as from the ZFMK biobank.

Conflict of interest disclosure

The authors declare that they comply with the PCI rule of having no financial conflicts of interest in relation to the content of the article.

Funding

This study belongs to the xHARES consortium funded by the R'Life initiative of the Academy of Finland, grant no. 329264.

References

Angerbjörn A, Schai-Braun SC (2022) Mountain Hare Lepus timidus Linnaeus, 1758. . In: *Handbook of the Mammals of Europe Handbook of the Mammals of Europe*, Hackländer K., Zachos F.E. (eds.) pp. 1-29. Springer: Cham

Bai Y, Lin W, Xu J, Song J, Yang D, Chen YE, Li L, Li Y, Wang Z, Zhang J (2021) Improving the genome assembly of rabbits with long-read sequencing. *Genomics* 113: 3216-3223

Beklemisheva VR, Romanenko SA, Biltueva LS, Trifonov VA, Vorobieva NV, Serdukova NA, Rubtsova NV, Brandler OV, O'Brien PC, Yang F *et al* (2011) Reconstruction of karyotype evolution in core Glires. I. The genome homology revealed by comparative chromosome painting. *Chromosome Res* 19: 549-565

Belaghzal H, Dekker J, Gibcus JH (2017) Hi-C 2.0: An optimized Hi-C procedure for high-resolution genome-wide mapping of chromosome conformation. *Methods* 123: 56-65

Birney E, Clamp M, Durbin R (2004) GeneWise and Genomewise. *Genome Res* 14: 988-995

Blaxter M, Archibald JM, Childers AK, Coddington JA, Crandall KA, Di Palma F, Durbin R, Edwards SV, Graves JAM, Hackett KJ *et al* (2022) Why sequence all eukaryotes? *P Natl Acad Sci USA* 119

Bruna T, Hoff KJ, Lomsadze A, Stanke M, Borodovsky M (2021) BRAKER2: automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database. *NAR Genom Bioinform* 3: lqaa108

Caravaggi A, Leach K, Santilli F, Rintala J, Helle P, Tiainen J, Bisi F, Martinoli A, Montgomery WI, Reid N (2017) Niche overlap of mountain hare subspecies and the vulnerability of their ranges to invasion by the European hare; the (bad) luck of the Irish. *Biol Invasions* 19: 655-674

Carneiro M, Rubin CJ, Di Palma F, Albert FW, Alfoldi J, Martinez Barrio A, Pielberg G, Rafati N, Sayyab S, Turner-Maier J *et al* (2014) Rabbit genome analysis reveals a polygenic basis for phenotypic change during domestication. *Science* 345: 1074-1079

506 Ceballos FC, Joshi PK, Clark DW, Ramsay M, Wilson JF (2018) Runs of homozygosity: windows into
507 population history and trait architecture. *Nat Rev Genet* 19: 220-234
508 Chen SF, Zhou YQ, Chen YR, Gu J (2018) fastp: an ultra-fast all-in-one FASTQ preprocessor.
509 *Bioinformatics* 34: 884-890
510 Cheng H, Concepcion GT, Feng X, Zhang H, Li H (2021) Haplotype-resolved de novo assembly using
511 phased assembly graphs with hifiasm. *Nat Methods* 18: 170-175
512 Crews S, Ojala D, Posakony J, Nishiguchi J, Attardi G (1979) Nucleotide sequence of a region of
513 human mitochondrial DNA containing the precisely identified origin of replication. *Nature*
514 277: 192-198
515 Fazekas J, Grunt TW, Jensen-Jarolim E, Singer J (2017) Long term storage in liquid nitrogen leads
516 to only minor phenotypic and gene expression changes in the mammary carcinoma model
517 cell line BT474. *Oncotarget* 8: 35076-35087
518 Ferreira MS, Jones MR, Callahan CM, Farelo L, Tolesa Z, Suchentrunk F, Boursot P, Mills LS, Alves
519 PC, Good JM *et al* (2021) The Legacy of Recurrent Introgression during the Radiation of
520 Hares. *Syst Biol* 70: 593-607
521 Forsyth NR, Elder FF, Shay JW, Wright WE (2005) Lagomorphs (rabbits, pikas and hares) do not
522 use telomere-directed replicative aging in vitro. *Mech Ageing Dev* 126: 685-691
523 Fuste JM, Wanrooij S, Jemt E, Granycome CE, Cluett TJ, Shi Y, Atanassova N, Holt IJ, Gustafsson
524 CM, Falkenberg M (2010) Mitochondrial RNA polymerase is needed for activation of the
525 origin of light-strand DNA replication. *Mol Cell* 37: 67-78
526 Gaertner K, Michell C, Tapanainen R, Goffart S, Saari S, Soininmaki M, Dufour E, Pohjoismaki JLO
527 (2022) Molecular phenotyping uncovers differences in basic housekeeping functions among
528 closely related species of hares (Lepus spp., Lagomorpha: Leporidae). *Mol Ecol*
529 Gertz EM, Yu YK, Agarwala R, Schaffer AA, Altschul SF (2006) Composition-based statistics and
530 translated nucleotide searches: improving the TBLASTN module of BLAST. *BMC Biol* 4: 41
531 Giska I, Farelo L, Pimenta J, Seixas FA, Ferreira MS, Marques JP, Miranda I, Letty J, Jenny H,
532 Hacklander K *et al* (2019) Introgression drives repeated evolution of winter coat color
533 polymorphism in hares. *Proc Natl Acad Sci U S A* 116: 24150-24156
534 Giska I, Pimenta J, Farelo L, Boursot P, Hacklander K, Jenny H, Reid N, Montgomery WI, Prodohl
535 PA, Alves PC *et al* (2022) The evolutionary pathways for local adaptation in mountain hares.
536 *Mol Ecol* 31: 1487-1503
537 Gustavsson I (1972) Mitotic and meiotic chromosomes of the variable hare (Lepus timidus L.), the
538 common hare (Lepus europaeus Pall.) and their hybrids. *Hereditas* 67: 27-34
539 Howe K, Chow W, Collins J, Pelan S, Pointon D-L, Sims Y, Torrance J, Tracey A, Wood J (2021)
540 Significantly improving the quality of genome assemblies through curation. *GigaScience* 10
541 Jansson G, Pehrson A (2007) The recent expansion of the brown hare (Lepus europaeus) in
542 Sweden with possible implications to the mountain hare (L. timidus). *Eur J Wildlife Res* 53:
543 125-130
544 Juhling F, Putz J, Bernt M, Donath A, Middendorf M, Florentz C, Stadler PF (2012) Improved
545 systematic tRNA gene annotation allows new insights into the evolution of mitochondrial
546 tRNA structures and into the mechanisms of mitochondrial genome rearrangements.
547 *Nucleic Acids Res* 40: 2833-2845
548 Juric I, Aeschbacher S, Coop G (2016) The Strength of Selection against Neanderthal Introgression.
549 *PLoS Genet* 12: e1006340

550 Kim D, Paggi JM, Park C, Bennett C, Salzberg SL (2019) Graph-based genome alignment and
551 genotyping with HISAT2 and HISAT-genotype. *Nat Biotechnol* 37: 907-915
552 Korstanje R, O'Brien PC, Yang F, Rens W, Bosma AA, van Lith HA, van Zutphen LF, Ferguson-Smith
553 MA (1999) Complete homology maps of the rabbit (Oryctolagus cuniculus) and human by
554 reciprocal chromosome painting. *Cytogenet Cell Genet* 86: 317-322
555 Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, Jones SJ, Marra MA (2009)
556 Circos: an information aesthetic for comparative genomics. *Genome Res* 19: 1639-1645
557 Lawniczak MKN, Durbin R, Flicek P, Lindblad-Toh K, Wei X, Archibald JM, Baker WJ, Belov K, Blaxter
558 ML, Marques Bonet T *et al* (2022) Standards recommendations for the Earth BioGenome
559 Project. *Proc Natl Acad Sci U S A* 119
560 Levanen R, Kunnasranta M, Pohjoismaki J (2018a) Mitochondrial DNA introgression at the
561 northern edge of the brown hare (Lepus europaeus) range. *Ann Zool Fenn* 55: 15-24
562 Levanen R, Thulin CG, Spong G, Pohjoismaki JLO (2018b) Widespread introgression of mountain
563 hare genes into Fennoscandian brown hare populations. *Plos One* 13
564 Li D, Liu CM, Luo R, Sadakane K, Lam TW (2015) MEGAHIT: an ultra-fast single-node solution for
565 large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* 31:
566 1674-1676
567 Li H (2018) Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* 34: 3094-
568 3100
569 Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform.
570 *Bioinformatics* 25: 1754-1760
571 Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, Amit I, Lajoie BR,
572 Sabo PJ, Dorschner MO *et al* (2009) Comprehensive mapping of long-range interactions
573 reveals folding principles of the human genome. *Science* 326: 289-293
574 Manni M, Berkeley MR, Seppey M, Zdobnov EM (2021) BUSCO: Assessing Genomic Data Quality
575 and Beyond. *Curr Protoc* 1: e323
576 Marques JP, Seixas FA, Farelo L, Callahan CM, Good JM, Montgomery WI, Reid N, Alves PC, Boursot
577 P, Melo-Ferreira J (2020) An Annotated Draft Genome of the Mountain Hare (Lepus timidus).
578 *Genome Biol Evol* 12: 3656-3662
579 Melo-Ferreira J, Alves PC, Freitas H, Ferrand N, Boursot P (2009) The genomic legacy from the
580 extinct Lepus timidus to the three hare species of Iberia: contrast between mtDNA, sex
581 chromosomes and autosomes. *Mol Ecol* 18: 2643-2658
582 Melo-Ferreira J, Vilela J, Fonseca MM, da Fonseca RR, Boursot P, Alves PC (2014) The elusive
583 nature of adaptive mitochondrial DNA evolution of an arctic lineage prone to frequent
584 introgression. *Genome Biol Evol* 6: 886-896
585 Meng G, Li Y, Yang C, Liu S (2019) MitoZ: a toolkit for animal mitochondrial genome assembly,
586 annotation and visualization. *Nucleic Acids Res* 47: e63
587 Michell CT, Pohjoismaki JLO, Spong G, Thulin CG (2022) Mountain- and brown hare genetic
588 polymorphisms to survey local adaptations and conservation status of the heath hare
589 (Lepus timidus sylvaticus, Nilsson 1831). *Sci Data* 9: 667
590 Mikheenko A, Prjibelski A, Saveliev V, Antipov D, Gurevich A (2018) Versatile genome assembly
591 evaluation with QUAST-LG. *Bioinformatics* 34: i142-i150
592 Open2C, Abdennur N, Fudenberg G, Flyamer IM, Galitsyna AA, Goloborodko A, Imakaev M, Venev
593 SV (2023) Pairtools: from sequencing data to chromosome contacts. *bioRxiv*

Ou S, Su W, Liao Y, Chougule K, Agda JRA, Hellinga AJ, Lugo CSB, Elliott TA, Ware D, Peterson T *et al* (2019) Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. *Genome Biol* 20: 275

Pohjoismaki JLO, Michell C, Levanen R, Smith S (2021) Hybridization with mountain hares increases the functional allelic repertoire in brown hares. *Sci Rep* 11: 15771

Rehnus M, Bollmann K, Schmatz DR, Hacklander K, Braunisch V (2018) Alpine glacial relict species losing out to climate change: The case of the fragmented mountain hare population (Lepus timidus) in the Alps. *Glob Chang Biol* 24: 3236-3253

Reid N (2011) European hare (Lepus europaeus) invasion ecology: implication for the conservation of the endemic Irish hare (Lepus timidus hibernicus). *Biol Invasions* 13: 559-569

Seluanov A, Vaidya A, Gorbunova V (2010) Establishing primary adult fibroblast cultures from rodents. *J Vis Exp*

Skov L, Coll Macia M, Lucotte EA, Cavassim MIA, Castellano D, Schierup MH, Munch K (2023) Extraordinary selection on the human X chromosome associated with archaic admixture. *Cell Genom* 3: 100274

Smit AFA, Hubley R, Green P, 2013-2015. RepeatMasker Open-4.0. http://www.repeatmasker.org.

Smith S, Sandoval-Castellanos E, Lagerholm VK, Napierala H, Sablin M, Von Seth J, Fladerer FA, Germonpre M, Wojtal P, Miller R *et al* (2017) Nonreceding hare lines: genetic continuity since the Late Pleistocene in European mountain hares (Lepus timidus). *Biol J Linn Soc* 120: 891-908

Thulin C-G (2003) The distribution of mountain hares (Lepus timidus, L. 1758) in Europe: A challenge from brown hares (L. europaeus, Pall. 1778)? *Mammal Rev* 33: 29-42

Thulin CG, Winiger A, Tallian AG, Kindberg J (2021) Hunting harvest data in Sweden indicate precipitous decline in the native mountain hare subspecies Lepus timidus sylvaticus (heath hare). *J Nat Conserv* 64

Vinogradov AE (1998) Genome size and GC-percent in vertebrates as determined by flow cytometry: The triangular relationship. *Cytometry* 31: 100-109

Wheeler TJ, Eddy SR (2013) nhmmer: DNA homology search with profile HMMs. *Bioinformatics* 29: 2487-2489

Wolf JB, Lindell J, Backstrom N (2010) Speciation genetics: current status and evolving approaches. *Philos Trans R Soc Lond B Biol Sci* 365: 1717-1733

Zhou C, McCarthy SA, Durbin R (2023) YaHS: yet another Hi-C scaffolding tool. *Bioinformatics* 39

Zimova M, Giery ST, Newey S, Nowak JJ, Spencer M, Mills LS (2020) Lack of phenological shift leads to increased camouflage mismatch in mountain hares. *Proc Biol Sci* 287: 20201786

Zimova M, Mills LS, Nowak JJ (2016) High fitness costs of climate change-induced camouflage mismatch. *Ecol Lett* 19: 299-307