Supplementary information

Genome shows no recent inbreeding in near-extinction woolly rhinoceros sample found in ancient wolf's stomach

S.M. Guðjónsdóttir, E. Lord *et al*.

List of supplementary figures

List of supplementary tables (attached separately as an Excel file)

List of supplementary texts

Figure S1. Tumat_14k specimen. Approximate size: 4 x 3 cm.

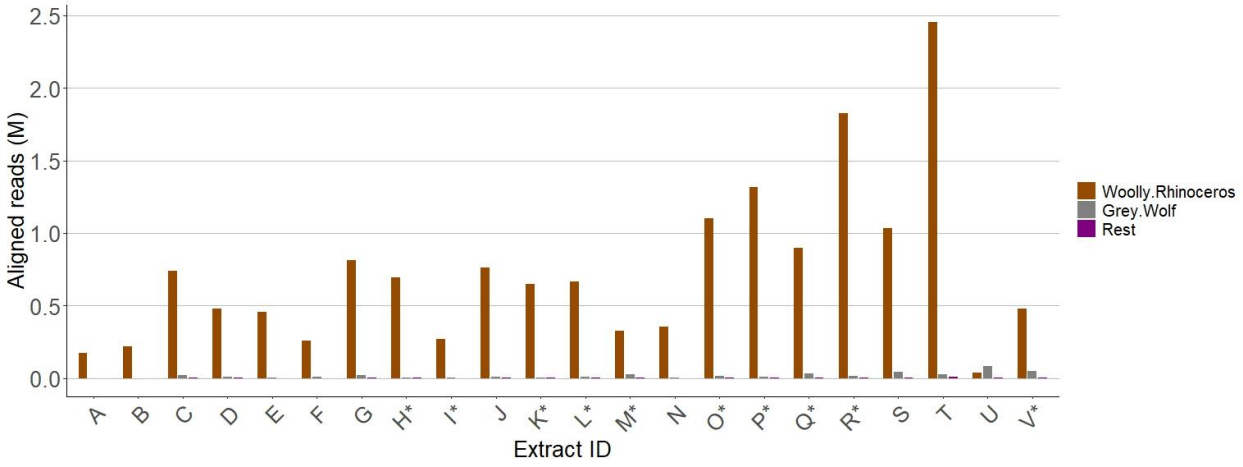


Figure S2. Presence of wolf DNA in the different Tumat_14k DNA extracts. The y axis shows the number of reads aligned to each reference mitogenome. For visualisation purposes, woolly rhinoceros and grey wolf are displayed separately while the rest (human, pig, cow, mouse and chicken) are merged into a single category. Extract U was excluded from all subsequent analyses. *Extracts used for the second round of sequencing.

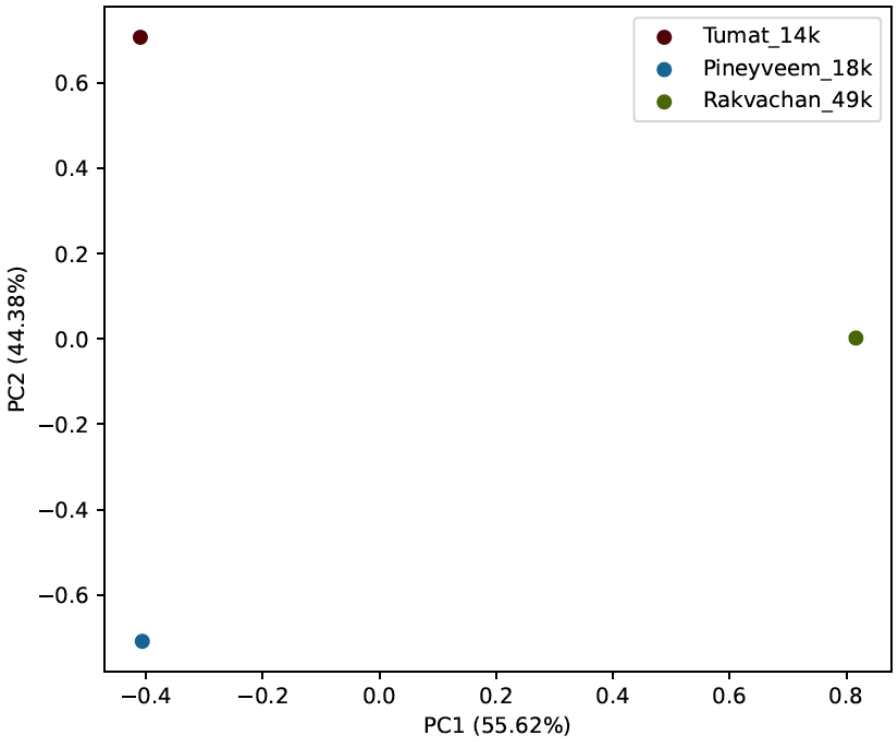
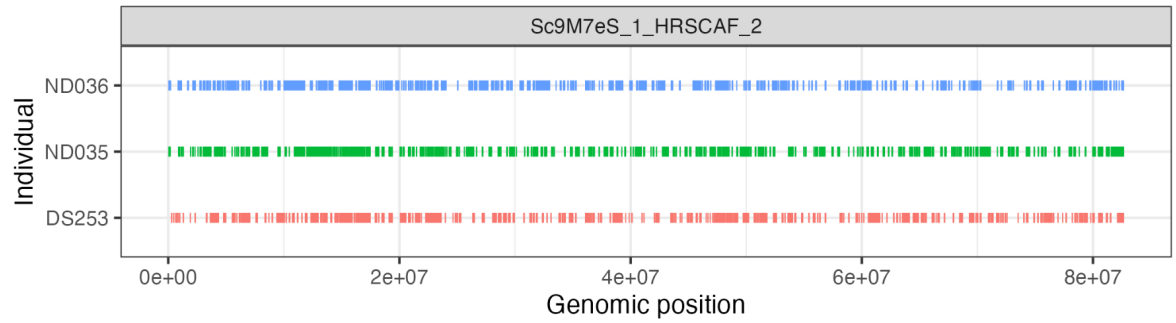


Figure S3. Principal component analysis using transversion-only dataset.

A)



B)

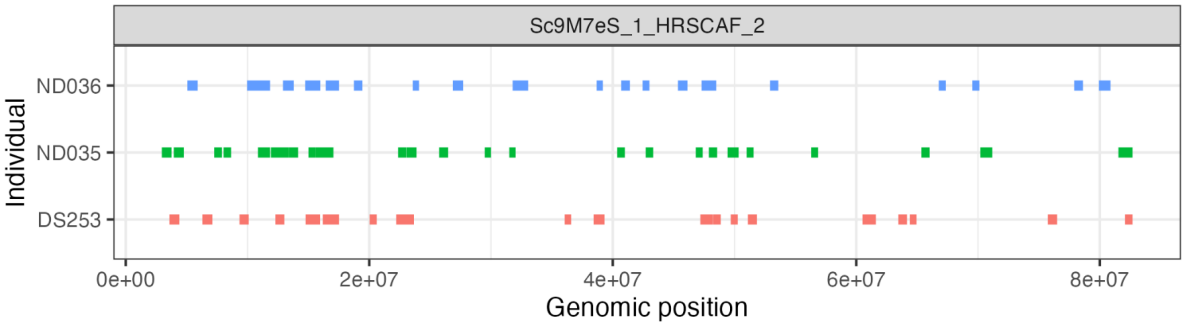


Figure S4. Distribution of Runs of Homozygosity (ROHs) **A)** above 100 kb and **B)** above 500 kb throughout scaffold Sc9M7eS_1_HRSCAF_2. ROHs were inferred using PLINK.

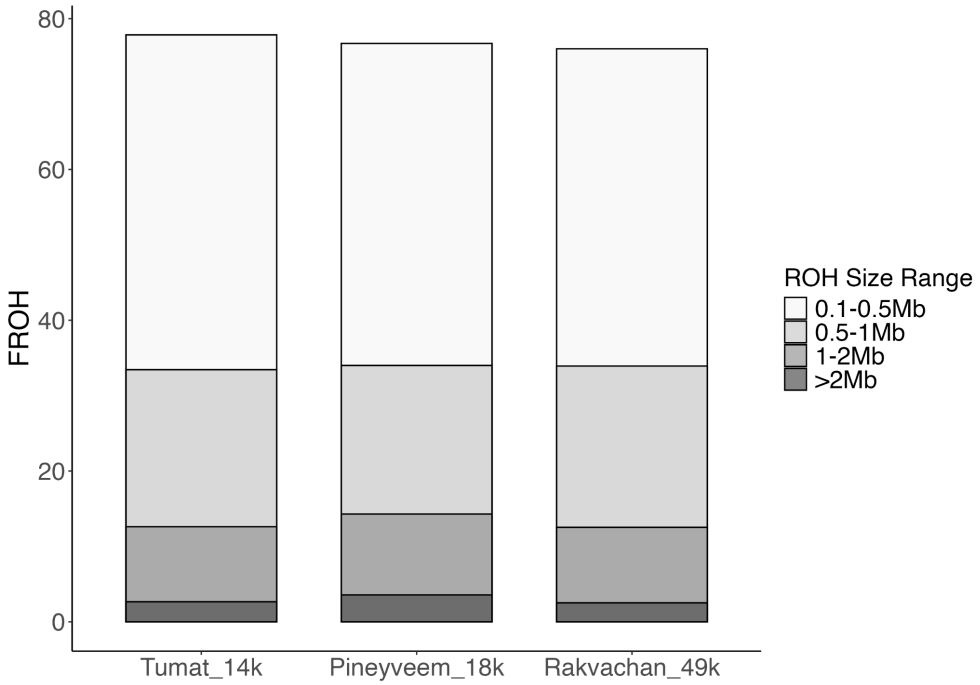


Figure S5. F_{ROH} estimates for all three samples (Tumat_14k, Pineyveem_18k and Rakvachan_49k) after removing transitions using BCFtools/RoH.

Text S1 - High-coverage shotgun sequencing

1.1 Generating a high coverage genome for Tumat_14k

The first round of sequencing of Tumat_14k (DS253) displayed considerable variation in DNA quality between the 20 different extractions (Table S1), with endogenous DNA content ranging from 1.9% to 8.3% and PCR duplicates from 15% to 42%. Endogenous content was estimated prior to duplicate removal as the percentage of total sequencing reads aligning to the reference genome. The low endogenous DNA content was expected due to the conditions in which the sample was preserved and as it had shown low levels on previous DNA extracts. We also assessed the genome recovery rate (GRR), which we defined as the portion of reads aligning to the reference genome after quality filtering steps, as well as removal of PCR duplicates and mitochondrial-linked scaffolds. Removing mitochondrial DNA to estimate the final GRR was important as mitogenomes occur in higher quantities than nuclear DNA within cells and could inflate the interpretation of endogenous content. The final portion of reads eventually used for downstream analyses (GRR) was below 5% for all the extracts after quality filtering steps.

Given the low GRR of the extracts, we made careful calculations for 10 extracts with the highest value to undergo additional library preparation and another round of sequencing to ensure enough high-quality endogenous DNA yield. We aimed for a 10X coverage genome to comparatively analyse it alongside two other high-coverage genomes. To achieve that, we first assigned a rough estimation of the depth of coverage we aimed to get from each of the 10 extracts from the second round of sequencing. Next, we calculated the number of reads required from each extract to reach that estimated coverage (Table 1) following the standard procedure used at the Centre for Palaeogenetics, demonstrated with equation 1 as

$$\frac{Reference\ genome\ length\ (bp)\ x\ desired\ coverage}{average\ read\ length\ (bp)\ x\ genome\ recovery\ rate} \quad (1).$$

In order to produce the number of reads needed from each extract, we prepared an additional library per extract and generated a total of 190 separate indexing PCRs (across 20 libraries) and estimated the concentration per sequencing library. When pooling together the libraries for the second round of sequencing, we estimated their final concentration in the pool by combining their estimated concentration with the relative contribution needed from each based on the number of reads desired. The amount of reads generated from the second round of sequencing for each extract was remarkably close to the estimated value.

Table 1: Overview of sequencing results for 20 of Tumat_14k extractions. Showing outcome from sequencing round 1 as total reads, complexity, number of reads from MQ30 uniq mapping to autosomes and genome recovery rate (final proportion of reads aligning to the reference genome after filtering steps; GRR). Desired number of reads from 10 extracts used in sequencing round 2 needed to reach desired coverage, compared with the total number of reads produced from the second round of sequencing.

Extract ID	Total reads from sequencing 1 (M)	Complexity (%)	Total MQ30 uniq reads mapped to autosomes (M)	Genome Recovery Rate (GRR %)	Reads for desired coverage (M)	Total reads from sequencing 2 (M)
C	96	65.47	0.9	0.85	-	-
D	71	67.92	0.6	0.91	-	-
E	65	85.08	1.1	1.67	-	-
F	64	73.28	0.5	0.81	-	-
G	77	64.36	1.3	1.69	-	-
H	87	80.71	2.7	3.05	1,380	1,286
I	57	81.11	1.4	2.39	432	380
J	104	73.55	2.0	1.93	-	-
K	92	65.01	2.3	2.47	419	396
L	63	72.22	2.2	3.47	1,453	940
M	63	69.81	1.5	2.36	432	542
N	55	81.30	1	1.81	-	-
O	99	79.21	3.9	3.92	1,519	1,484
P	110	57.61	2.6	2.36	439	339
Q	79	75.53	3.5	4.4	2,422	2,562
R	150	74.33	4.7	3.15	1,277	1,257
S	104	61.99	0.9	0.83	-	-
T	130	60.71	1.6	1.26	-	-
U	72	67.75	0.7	0.96	-	-
V	113	71.86	3.5	3.06	1,452	1,874

1.2 Samples & sequencing results

By combining all sequencing for Tumat_14k, the sample reached an average of 10.1X, comparable with the other two samples: 11X for Pineyveem_18k and 11.1X for Rakvachan_49k (Table S2). The coverage for the latter two is lower than reported in their original publications (Lord et al. 2020; Liu et al. 2021), likely because a more stringent quality filtering was applied in this analysis, including a second round of duplicate removal. The final endogenous DNA content for Tumat_14k was 5%, low compared to Pineyveem_18k and Rakvachan_49k with 56% and 35%, respectively. The overall number of duplicates was nearly twice as high for the stomach rhino sample. Aiming for higher sequencing depth could come at the cost of reduced complexity as unique fragments can get exhausted from the sample (Dehasque et al. 2022).

Text S2 - Assessment of wolf DNA in Tumat_14k

Since Tumat_14k was found in the stomach of a grey wolf, it was essential to assess the extent of wolf DNA in the sequencing data. Ancient samples often contain other DNA sources as well, so we used a competitive mapping (alignment) approach (Feuerborn et al. 2020) where the original merged reads were aligned to multiple mitogenomes. We created a concatenated fasta file using the reference mitogenomes of the woolly rhinoceros (*Coelodonta antiquitatis*, NC_012681.1), grey wolf (*Canis lupus*, NC_008092.1), human (*Homo sapiens*, NC_012920.1), pig (*Sus scorfa*, NC_000845.1), cow (*Bos taurus*, NC_006853.1), mouse (*Mus musculus*, NC_005089.1) and chicken (*Gallus gallus*, NC_001323.1). Each extract was aligned using the same approach described in the main text's methods section "Alignment of sequencing data". The results were analysed using SAMtools v1.17 idxstats which gave the number of reads that aligned to each mitogenome. For the competitive mapping, we used the reads from sequencing round 1 for extracts C-V as well as the fastq file from the published extract A (Lord et al. 2020) and a subsampled fastq file for extract B.

Most of the sample's extracts had minimal wolf DNA with under 5% of the aligned reads corresponding to the grey wolf mitogenome and under 0.4% to the rest of possible DNA sources, including humans (Fig S2). However, extract U displayed a high amount of wolf DNA, with 66% of reads aligning to the grey wolf mitogenome and only 31% to the woolly rhinoceros mitogenome. This extract was subsequently excluded from all downstream analyses to reduce the risk of wolf DNA bias in the dataset. As the analyses were conducted on a high-coverage genome, the chance of calling false SNPs (single nucleotide polymorphisms) due to wolf DNA bias in the dataset is minimal (Llamas et al. 2017; Renaud et al. 2019).

Nonetheless, since the mitogenome-based estimates potentially only offer a lower bound for the estimation of the amount of contamination present on the sequencing data and to corroborate that this contamination does not affect our inferences, we also performed a competitive mapping using the entire reference genomes for the target species (Sumatran rhinoceros) and the main source of contamination (grey wolf). We followed exactly the same procedures described in the main text (see methods sections "Data processing" and "Variant calling"), with the only difference that we excluded all reads mapping to grey wolf prior to variant calling.

Across the 20 extracts, an average of ~1% of the sequenced reads mapped to the grey wolf reference genome. After removing PCR duplicates and filtering for mapping quality 25, only ~0.03% of the sequenced reads aligned to grey wolf. Table 1 (below) shows the contamination estimations for each extract as estimated from the same screening round described in Text S1. We additionally included a library from ND036 (Rakvachan_49k) to compare the base levels of DNA mapping to grey wolf in a sample that theoretically should not contain any contamination of this kind, providing a baseline.

Table 1. Contamination estimates obtained from the competitive mapping approach using the concatenated sumatran rhinoceros and grey wolf reference genomes

Extract ID	% of total reads mapped to grey wolf	% of total reads mapped to grey wolf after MQ25
A	6.332	0.009
B	0.217	0.01
C	0.564	0.031
D	0.386	0.022
E	0.26	0.017
F	0.501	0.029
G	1.266	0.043
H	0.606	0.018
I	0.661	0.021
J	0.774	0.019
K	1.593	0.019
L	1.529	0.029
M	1.13	0.06
N	0.507	0.02
O	0.864	0.029
P	2.155	0.025
Q	1.388	0.059
R	0.714	0.021
S	0.733	0.053
T	0.636	0.03
U	0.743	0.131
V	1.018	0.058
ND036_08_L1	0.237	0.021

After variant calling, we obtained a genome coverage of 9.9x for Tumat_14k, almost identical to the one originally obtained. To corroborate that our estimations hold regardless of the alignment method used (from now on regarded as “non-competitive” and “competitive”) we subsampled the non-competitive based Tumat_14k genome from 10.1x to 9.9x and estimated genome-wide heterozygosity using direct counts from BCFtools (following the same procedures described in the main methods section. We got an estimated ~1.2 SNPs per 1,000bp for both approaches (including all types of variants). This demonstrates that wolf contamination does not have an effect on variant calling and downstream analyses.

Text S3 - Metagenomic screening

To assess the presence of ancient host-associated microbes and pathogens, we performed a metagenomic screening with the first module of the aMeta pipeline (commit 16554c6)(Pochon et al. 2023). In summary, it performs a quality control check with FastQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/), before and after adapter removal with Cutadapt(Martin 2011). Then it runs KrakenUniq(Breitwieser et al. 2018) to classify the reads using the aMeta Microbial NCBI NT database, available on https://doi.org/10.17044/scilifelab.20518251. This database comprises all microbial genomic information from NCBI NT (viruses, bacteria, archaea but also eukaryotic microbes like fungi, protozoa and parasitic worms) and a few other eukaryotic complete genomes. Subsequently, KrakenUniq outputs are filtered to keep species that have at least 200 reads and 1,000 *k*-mers.

In parallel, aMeta performs an alignment to a microbial-human genome database https://doi.org/10.17044/scilifelab.21185887 using Bowtie2(Langmead & Salzberg 2012) and it creates deamination plots for the authentication of the microbial ancient status with mapDamage2(Jónsson et al. 2013). Since two out of three samples were USER-treated, we also used PMDtools(Skoglund et al. 2014) to estimate post-mortem damage based on CpG sites.

Interestingly, we could not identify any ancient microbe in the untreated rhino sample. Furthermore, investigation of a deamination profile based on CpG sites for USER-treated samples did not reveal ancient microbial organisms either, maybe due to lack of coverage, but the presence of mostly modern contaminants cannot be excluded (Table S4).

Several organisms were found in two out of three samples but are interpreted as environmental contamination like *Cupriavidus metallidurans*, *Cutibacterium acnes*, *Enterococcus faecalis*, *Herbaspirillum seropedicae*, *Paeniclostridium sordellii*, *Rhodopseudomonas palustris*, *Staphylococcus epidermidis*, *Streptococcus canis* and *Variovorax paradoxus*. Although generally considered an opportunistic pathogen in dogs, *S. canis* was identified in both DS253 and ND035 making it more likely to be a sign of contamination. Additionally, *Collimonas spp.*, *Dictyostelia spp.* (amoeba), *Jonesia denitrificans*, *Sphingomonas melonis* and *Pseudomonas yamanorum* are associated with soil. Furthermore, *Clostridia spp.*, *Listeria monocytogenes* and *Paraclostridium bifermentans* are associated with the intestinal tract of animals but are also commonly found in soils. Moreover, *Streptococcus pyogenes* might be due to human contamination. Finally, the *Carnobacteria* and *Lactobacilli* species found are generally associated with meat kept in cold environments.

Text S4 - Additional analyses without removing transitions

Demographic analyses

PSMC was performed with and without transitions (Figure 1). For both analyses we also removed the last 10,000 years (6 steps).

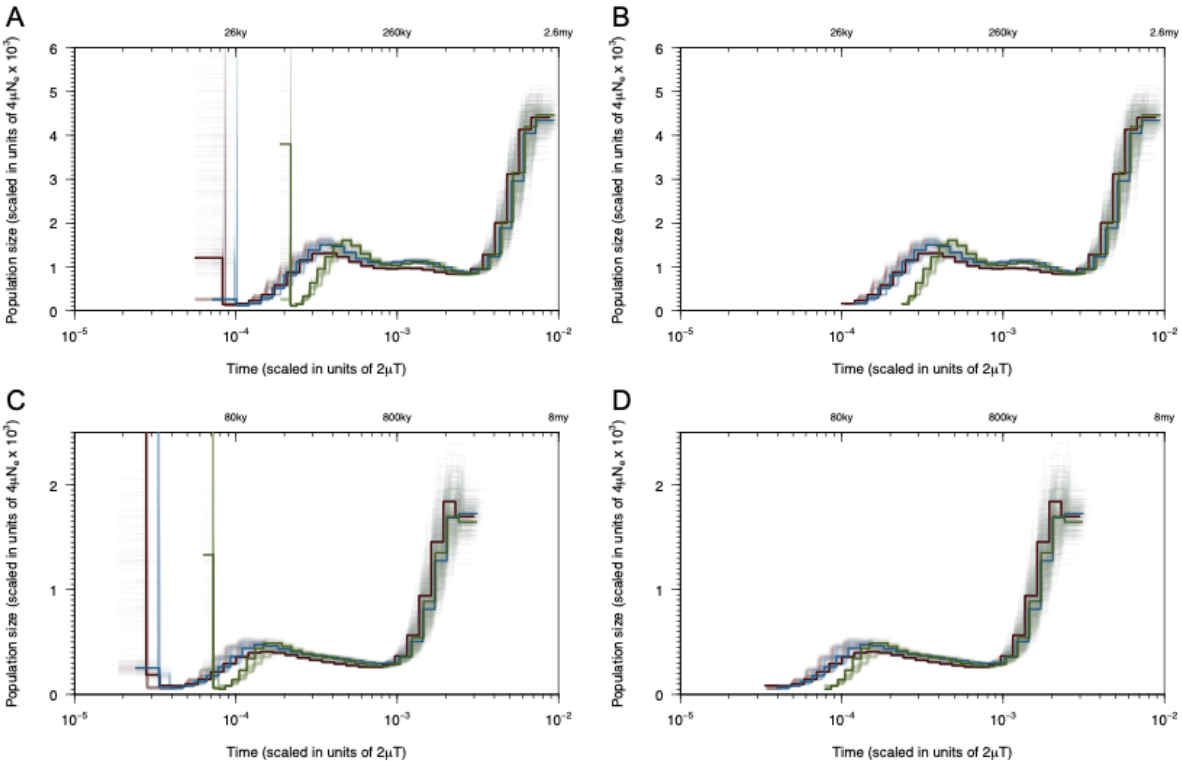


Figure 1: All plots show DS253 in red, ND035 in blue, and ND036 in green, with 300 bootstrap replicates in pale shades. **A** shows the PSMC including transitions, scaled using a mutation rate of 2.34x10-8 substitutions per site per generation and a generation time of 12 years. **B** shows the PSMC including transitions, scaled using a mutation rate of 2.34x10-8 substitutions per site per generation and a generation time of 12 years with the most recent 10000 years removed. **C** shows the PSMC excluding transitions, scaled using a mutation rate of 0.78x10-8 substitutions per site per generation and a generation time of 12 years. **D** shows the PSMC including transitions, scaled using a mutation rate of 0.78x10-8 substitutions per site per generation and a generation time of 12 years with the most recent 10,000 years removed.

Heterozygosity and inbreeding

The downstream analyses were also performed using with and without transitions. For the two younger samples the estimated population mutation rate (θ) was 1.77 SNPs per 1,000bp (95% CI: 1.77-1.77) for Tumat_14k and 1.64 (95% CI: 1.64-1.65) for Pineyveem_18k. Additionally we estimated genome-wide heterozygosity using allele counts from variant calling which revealed ~1.2 SNPs per 1,000bp for both samples. These two estimates differ due to different approaches in estimating heterozygosity, with θ accounting for possible sequencing error rate and recombination within the genome(Haubold et al. 2010). The estimates for the younger two samples decreased

after removing transitions from the genome, as they rely solely on heterozygous sites that occur from transversions (around ⅓ of total SNPs).

By including transitions in inbreeding estimations, Tumat_14k and Pineyveem_18k had 41% and 42% of their genome within homozygous segments, respectively. The longest ROH segment was 5.2 Mb for Tumat_14k and 4.7 for Pineyveem_18k. For both samples, 98% of all ROH windows were under 1 Mb long.

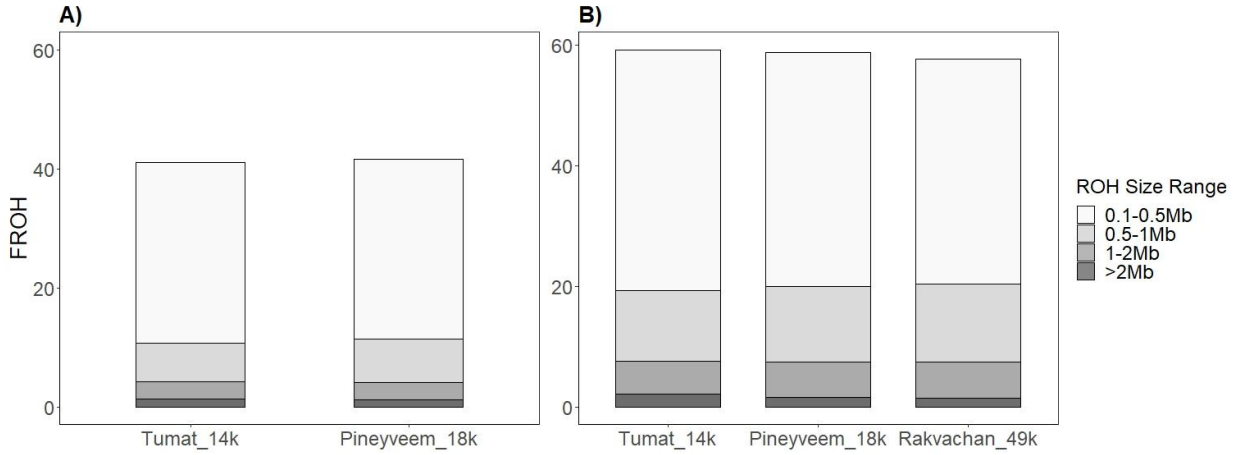


Figure 2: A) F_{ROH} results for the two younger samples with transitions/transversions B) F_{ROH} for all three samples (Tumat_14k, Pineyveem_18k and Rakvachan_49k) after removing transitions. White and grey coloured markings show the F_{ROH} value for short segments while black indicates long ROH window sizes >2Mb.

Statistical tests

We tested for differences in the average size of ROHs among the samples for four different size thresholds >0.1Mb, >0.5Mb, >1Mb and >2Mb. As the ROH window sizes followed a non-normal frequency distribution, non-parametric statistical tests were used. We used the Wilcoxon-rank-sum U test for the two younger samples including transitions (Table 2) and the Kruskal-Wallis test for all three samples with transversions only (Table S3). The null hypotheses H_0 in both cases were that the frequency distributions of ROH sizes were equal between samples and were conducted using R v.4.2.3(R Core Team 2023). As the tests were being applied simultaneously on a single dataset, the Bonferroni Correction was applied to avoid generating false-positives(Dunn 1961). With a 99% confidence interval, the critical p-value was set as 0.01, and as we were analysing four different size thresholds for two types of statistical tests simultaneously, the Bonferroni Correction set the critical p-value to 0.01/8 = 0.001 (Table S3).

Table 2: Summary results comparing different sizes of Runs of Homozygosity (ROH).

ROH length threshold	**> 0.1Mb**	**>0.5Mb**	**>1Mb**	**>2Mb**
DS253-ND035 **Wilcoxon U test (p-value)**	0.031	0.635	0.405	0.173
DS253				
Nr. of ROHs	3,850	288	60	8
Mean length (Mb)	0.25	0.86	1.62	3.75
Median length (Mb)	0.17	0.69	1.29	4.01
ND035				
Nr. of ROHs	3,820	314	65	10
Mean length (Mb)	0.25	0.84	1.48	2.8
Median length (Mb)	0.18	0.7	1.23	2.64

References

Breitwieser FP, Baker DN, Salzberg SL. 2018. KrakenUniq: confident and fast metagenomics classification using unique k-mer counts. Genome Biol. 19:198.

Dehasque M et al. 2022. Development and Optimization of a Silica Column-Based Extraction Protocol for Ancient DNA. Genes . 13. doi: 10.3390/genes13040687.

Dunn OJ. 1961. MULTIPLE COMPARISONS AMONG MEANS. J. Am. Stat. Assoc. 56:52–&.

Feuerborn TR et al. 2020. Competitive mapping allows for the identification and exclusion of human DNA contamination in ancient faunal genomic datasets. BMC Genomics. 21. doi: 10.1186/s12864-020-07229-y.

Haubold B, Pfaffelhuber P, Lynch M. 2010. mlRho - a program for estimating the population mutation and recombination rates from shotgun-sequenced diploid genomes. Mol. Ecol. 19:277–284.

Jónsson H, Ginolhac A, Schubert M, Johnson PLF, Orlando L. 2013. mapDamage2.0: fast approximate Bayesian estimates of ancient DNA damage parameters. Bioinformatics. 29:1682–1684.

Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. Nat. Methods. 9:357–359.

Liu SL et al. 2021. Ancient and modem genomes unravel the evolutionary history of the rhinoceros family. Cell. 184:4874–+.

Llamas B et al. 2017. From the field to the laboratory: Controlling DNA contamination in human ancient DNA research in the high-throughput sequencing era. STAR: Science & Technology of Archaeological Research. 3:1–14.

Lord E et al. 2020. Pre-extinction Demographic Stability and Genomic Signatures of Adaptation in the Woolly Rhinoceros. Curr. Biol. 30:3871–+.

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

Pochon Z et al. 2023. aMeta: an accurate and memory-efficient ancient metagenomic profiling workflow. Genome Biol. 24:242.

R Core Team. 2023. R: A language and environment for statistical computing. https://www.R-project.org/.

Renaud G, Hanghoj K, Korneliussen TS, Willerslev E, Orlando L. 2019. Joint Estimates of Heterozygosity and Runs of Homozygosity for Modern and Ancient Samples. Genetics. 212:587–614.

Skoglund P et al. 2014. Separating endogenous ancient DNA from modern day contamination in a Siberian Neandertal. Proc. Natl. Acad. Sci. U. S. A. 111:2229–2234.

Table S1 - Sequencing results for each tissue extract taken from Tumat_14k (DS253)

Extract ID	Total reads	Endogenous	Complexity (1-
A	22	10.97	19.74
B	5805	2.28	21.89
C	96	2.33	65.47
D	71	2.34	67.92
E	65	2.87	85.08
F	64	1.87	73.28
G	77	4.29	64.36
H*	1373	4.9	74.38
I*	437	3.99	77.52
J	104	3.89	73.55
K*	488	5.28	63.70
L*	1003	6.57	66.51
M*	605	5.11	64.52
N	55	3.26	81.30
O*	1583	6.53	73.01
P*	449	5.41	62.61
Q*	2641	8.23	66.51
R*	1407	5.87	72.09
S	104	2.48	61.99
T	130	3.46	60.71
U**	72	2.43	67.75
V*	1987	6.38	66.90

*Extracts used for second round of sequencing
**Extract excluded from analyses

Table S2 - Sample information and sequencing results.

Sample ID	**DS253**	**ND035**
Manuscript ID	Tumat_14k	Pineyveem_18k
Total reads (M)	18,638	896
Final reads aligned (M)	407	338
Endogenous %	5	56
Total duplicates %	40.1	17
Transition/transversion ratio	1.77	1.76
Depth	10.1X	11X
Breadth of coverage - sites >5x	0.801	0.841
C14 date	12,355±31 BP	15,260±65 BP
Calibrated median	14,393±202 cal BP	18,452±134 cal BP
C14 accession	ETH-99775	OxA-36568
Source locality	Tumat (Stomach content of	Pineyveem River,
Accession ID	SAMEA6246875	SAMEA6246871

ND036
Rakvachan_49k
2,395
474
35
25.1
1.69
11.1X
0.827
46,200±230 BP
48,497±434 cal BP
OxA-36569
Rakvachan River, North
SAMN17167289

Table S3 - Number of derived alleles out of 17888 derived sites identified in Lord et al. 2020

	transitions included	transitions removed
Sample	**Number of**	**Number of positions**
Tumat_14k (DS253)	11405	4656
Pineyveem_18k (ND035)	11942	4841
Rakvachan_49k (ND036)	11092	4431
Proportion derived	**Number of**	**Number of positions**
3/3 derived	9762	3979
2/3 derived	1895	767
1/3 derived	986	422

Table S3 - Krakenuniq statistics averaged per fastq containing species with at lea

Tumat_14k (DS253)

Species	TaxID	FastQ	Avg	Avg	mers
Cutibacterium acnes	1747	2	238	236	7,313
Carnobacterium maltaromaticum	2751	20	13,133	8,887	383,527
Jonesia denitrificans	43674	1	250	250	7,034
Carnobacterium divergens	2748	1	65,711	236	1,265,689
Streptococcus canis	1329	3	569	500	8,423
Lactobacillus sakei	1599	20	95,023	32,642	1,093,411
Streptococcus pyogenes	1314	1	265	226	3,651
Sphingomonas melonis	152682	1	311	265	3,967
Clostridium novyi	1542	20	143,938	456	974,638
Staphylococcus epidermidis	1282	1	253	204	1,810
Lactobacillus curvatus	28038	20	5,009	2,366	28,476
Clostridium septicum	1504	20	5,286	4,787	24,308
Clostridiaceae bacterium 14S(	2E+06	20	744,190	6,161	1,735,309
Dictyostelium discoideum	44689	20	3,475	922	16,147
Paeniclostridium sordellii	1505	20	881,904	527,702	2,402,353
Clostridium perfringens	1502	20	14,817	7,059	67,699
Listeria monocytogenes	1639	1	328	262	1,157
Dictyostelium purpureum	5786	9	1,797	305	4,528
Enterococcus faecalis	1351	5	1,156	283	2,997
Collimonas pratensis	279113	20	3,141	2,016	4,951
Variovorax paradoxus	34073	1	3,515	222	5,640
Paraclostridium bifermentans	1490	10	4,319	393	6,026
Collimonas fungivorans	158899	20	4,703	1,188	5,753
Rhodopseudomonas palustris	1076	19	2,787	1,330	3,488
Pseudomonas yamanorum	515393	20	45,720	28,950	55,035
Collimonas arenae	279058	20	5,083	3,119	4,945
Clostridium botulinum	1491	20	85,276	30,034	68,355
Cupriavidus metallidurans	119219	1	971	703	1,004
Herbaspirillum seropedicae	964	20	3,267	2,810	2,818

Pineyveem_18k (ND035)

Species	TaxID	FastQ	Avg	Avg	mers
Staphylococcus epidermidis	1282	1	711	614	29631
Staphylococcus hominis	1290	1	1070	970	41022
Cutibacterium acnes	1747	1	1028	1009	39187
Staphylococcus saprophyticus	29385	1	366	321	12514
Malassezia globosa	76773	1	332	208	8861
Serratia fonticola	47917	1	2373	550	37694
Pseudomonas arsenicoxydans	702115	1	3963	1146	56802
Micrococcus luteus	1270	1	2772	1678	35453

Streptococcus canis	1329	1	540	477	6248
Rhodococcus qingshengii	334542	1	1969	641	17947
Rhodococcus erythropolis	1833	1	8202	2584	74310
Rhodococcus fascians	1828	1	1199	283	9519
Aminobacter sp. MSH1	374606	1	860	776	5621
Mesorhizobium japonicum	2E+06	1	1125	847	7274
Rhodopseudomonas palustris	1076	1	10756	2752	64937
Brevundimonas naejangsanensis	588932	1	2520	624	14208
Bradyrhizobium erythrophlei	1E+06	1	18250	497	102278
Pseudomonas aeruginosa	287	1	538	311	2894
Pseudomonas fluorescens	294	1	2068	318	11100
Propionibacterium freudenreichii	1744	1	616	542	3256
Rhizobacter gummiphilus	946333	1	1196	1196	6214
Magnetospirillum gryphiswaldens	55518	1	413	319	2030
Oligotropha carboxidovorans	40137	1	905	880	4421
Sinorhizobium meliloti	382	1	600	461	2930
Caulobacter segnis	88688	1	343	343	1667
Caulobacter vibrioides	155892	1	663	443	3121
Pelagibacterium halotolerans	531813	1	530	530	2488
Sphingopyxis macrogoltabida	33050	1	374	220	1747
Gemmata obscuriglobus	114	1	397	396	1841
Lysobacter enzymogenes	69	1	736	386	3377
Methanosarcina barkeri	2208	1	2890	1736	13241
Pseudomonas stutzeri	316	1	1105	405	5038
Bradyrhizobium japonicum	375	1	1155	611	5233
Bradyrhizobium symbiodeficiens	1E+06	1	1198	741	5400
Corynebacterium imitans	156978	1	239	239	1077
Pseudomonas chlororaphis	587753	1	687	278	3092
Rhizobium leguminosarum	384	1	2559	996	11478
Brevibacterium aurantiacum	273384	1	359	273	1599
Ensifer adhaerens	106592	1	987	431	4367
Blastochloris viridis	1079	1	664	664	2937
Bradyrhizobium diazoefficiens	1E+06	1	710	467	3137
Cupriavidus metallidurans	119219	1	346	280	1502
Variovorax paradoxus	34073	1	4767	653	20643
Stenotrophomonas maltophilia	40324	1	1015	451	4356
Sphingomonas wittichii	160791	1	465	325	1990
Agrobacterium tumefaciens	358	1	1470	559	6285
Hydrogenophaga sp. PBL-H3	434010	1	828	828	3532
Pannonibacter phragmitetus	121719	1	596	309	2538
Kocuria rosea	1275	1	1363	680	5798
Sinorhizobium fredii	380	1	1261	456	5349
Methanosarcina mazei	2209	1	1293	742	5475
Cupriavidus taiwanensis	164546	1	677	408	2837

Lysobacter antibioticus	84531	1	462	271	1934
Rhodospirillum rubrum	1085	1	271	270	1125
Herbaspirillum seropedicae	964	1	259	203	1075
Achromobacter denitrificans	32002	1	330	220	1359
Rhodobacter sphaeroides	1063	1	771	525	3113
Acidovorax carolinensis	553814	1	2417	1622	9518
Methanosarcina siciliae	38027	1	1093	515	4217
Pseudomonas putida	303	1	1027	269	3951
Burkholderia gladioli	28095	1	480	357	1834
Rhizobium phaseoli	396	1	319	206	1220
Bordetella bronchialis	463025	1	281	241	1059
Achromobacter xylosoxidans	85698	1	916	236	3439
Alicycliphilus denitrificans	179636	1	1126	993	4214
Pandoraea thiooxydans	445709	1	273	273	1014
Mycobacteroides abscessus	36809	1	345	294	1253
Sorangium cellulosum	56	1	2807	832	10147
Azospirillum brasilense	192	1	924	412	3317
Burkholderia multivorans	87883	1	415	263	1470
Delftia tsuruhatensis	180282	1	524	243	1824
Ochrobactrum anthropi	529	1	351	288	1220
Acidovorax citrulli	80869	1	420	387	1436
Mycobacterium shigaense	722731	1	398	367	1302
Microterricola viridarii	412690	1	2195	274	7039
Mycolicibacterium gilvum	1804	1	890	664	2751
Mycolicibacterium aurum	1791	1	1956	1034	6020
Mycolicibacter terrae	1788	1	389	389	1194
Gordonia bronchialis	2054	1	335	335	1004
Mycolicibacterium smegmatis	1772	1	1086	869	3221
Ralstonia solanacearum	305	1	708	387	2085
Mycolicibacterium chitae	1792	1	967	966	2795
Mycobacterium avium	1764	1	464	239	1331
Clavibacter michiganensis	28447	1	1871	502	5349
Rhodococcus hoagii	43767	1	692	588	1943
Mycobacterium kansasii	1768	1	454	403	1272
Rhodococcus rhodochrous	1829	1	671	201	1851
Acidipropionibacterium jensenii	1749	1	502	399	1287
Tsukamurella tyrosinosolvens	57704	1	534	407	1310
Rathayibacter festucae	110937	1	557	271	1344
Acidipropionibacterium acidiprop	1748	1	473	414	1081
Cellulomonas fimi	1708	1	1224	1215	2759
Gordonia terrae	2055	1	525	386	1131
Streptomyces avermitilis	33903	1	489	385	1040
Intrasporangium calvum	53358	1	3196	2531	6794
Nocardia cyriacigeorgica	135487	1	1134	356	2400

Nocardia farcinica	37329	1	763	671	1589
Rhodococcus ruber	1830	1	842	658	1672
Nocardia brasiliensis	37326	1	1671	319	3257
Nocardia seriolae	37332	1	563	552	1098
Streptomyces albus	1888	1	1160	616	2057
Amycolatopsis keratiniphila	129921	1	740	228	1249
Nocardia terpenica	455432	1	1217	373	2026
Nocardiopsis dassonvillei	2014	1	1018	855	1664
Streptomyces venezuelae	54571	1	3155	1064	5137
Streptomyces rimosus	1927	1	658	637	1007
Streptomyces griseorubiginosus	67304	1	975	295	1400
Streptomyces lincolnensis	1915	1	709	636	1015
Actinoplanes friuliensis	196914	1	103389	478	142132
Amycolatopsis mediterranei	33910	1	2395	2395	3212

Rakvachan_49k (ND036)

Species	**TaxID**	**FastQ**	**Avg**	**Avg**	**Avg K-mers**
Cutibacterium acnes	1747	1	575	567	21773
Paeniclostridium sordellii	1505	1	574	405	21531
Pantoea agglomerans	549	1	1159	1002	30530
Kocuria rosea	1275	1	7446	3399	188166
Brevibacterium aurantiacum	273384	1	380	292	9225
Enterococcus faecalis	1351	1	502	472	12015
Exiguobacterium sp. N4-1P	2E+06	1	1921	317	44353
Rhodococcus fascians	1828	1	1146	308	20806
Moraxella osloensis	34062	1	462	217	7876
Psychrobacter cryohalolentis	330922	1	857	855	7616
Variovorax paradoxus	34073	1	5577	780	47999
Rhodopseudomonas palustris	1076	1	1428	353	10483
Mesorhizobium japonicum	2E+06	1	248	203	1749
Acidovorax carolinensis	553814	1	1056	716	6295
Rhizobacter gummiphilus	946333	1	667	667	3719
Alicycliphilus denitrificans	179636	1	420	369	2092
Hydrogenophaga sp. PBL-H3	434010	1	484	484	2341
Rhodobacter sphaeroides	1063	1	571	385	1904

ıst 200 reads, 1000 *k*-mers and a ratio of *k*-mers/reads >= 1

Avg	Avg Coverage	ratio	Also found in
1	0.0017	30.66	ND035, ND036
1	0.0708	30.13	
1	0.0026	28.14	
2	0.5241	19.26	
1	0.0042	15.02	ND035
3	0.1877	14	
1	0.0006	13.78	
1	0.0014	12.76	
5	0.2658	7.97	
2	0.0002	7.15	ND035
4	0.0063	7.09	
4	0.0086	5.5	
23	0.6972	4.87	
2	0.0011	4.78	
20	0.51	4.17	ND036
7	0.0046	4.05	
3	0.0001	3.53	
5	0.0003	2.52	
6	0.0003	2.5	ND036
8	0.0007	1.81	
7	0.0002	1.6	ND035, ND036
16	0.0021	1.49	
9	0.0007	1.4	
10	0.0001	1.29	ND035, ND036
8	0.0108	1.22	
14	0.0005	1.11	
21	0.0023	1.06	
14	0.0001	1.03	ND035
11	0.0004	1	ND035

Avg	Avg Coverage	ratio	Also found in
1	0.0041	41.68	DS253
1	0.0097	38.34	
1	0.0091	38.12	DS253, ND036
1	0.0036	34.19	
1	0.001	26.69	
1	0.0029	15.88	
1	0.0073	14.33	
1	0.0061	12.79	

1	0.0031	11.57	DS253
2	0.01	9.11	
2	0.0131	9.06	
1	0.0008	7.94	ND036
1	0.001	6.54	
1	0.0012	6.47	
2	0.0017	6.04	DS253, ND036
1	0.0027	5.64	
2	0.0044	5.6	
2	0.0001	5.38	
1	0.0001	5.37	
2	0.0006	5.29	
2	0.0011	5.2	
1	0.0004	4.92	
2	0.0012	4.89	
1	0.0002	4.88	
1	0.0004	4.86	
1	0.0004	4.71	
1	0.0007	4.69	
1	0.0002	4.67	
1	0.0002	4.64	
1	0.0004	4.59	
2	0.0011	4.58	
2	0.0001	4.56	
1	0.0006	4.53	
1	0.0007	4.51	
2	0.0005	4.51	
2	0.0001	4.5	
1	0.0002	4.49	
2	0.0002	4.45	ND036
1	0.0003	4.42	
2	0.0009	4.42	
2	0.0005	4.42	
2	0.0001	4.34	DS253
2	0.0007	4.33	DS253, ND036
1	0.0001	4.29	
1	0.0003	4.28	
2	0.0002	4.28	
2	0.0009	4.27	
2	0.0003	4.26	
2	0.0008	4.25	ND036
1	0.0002	4.24	
2	0.0008	4.23	
1	0.0001	4.19	

2	0.0003	4.19	
2	0.0003	4.15	
2	0.0001	4.15	DS253
2	0.0002	4.12	
2	0.0002	4.04	
2	0.0015	3.94	
2	0.0007	3.86	
2	0.0001	3.85	
2	0.0001	3.82	
2	0.0001	3.82	
2	0.0002	3.77	
2	0.0001	3.75	
2	0.0008	3.74	
2	0.0002	3.71	
2	0.0001	3.63	
2	0.0002	3.61	
2	0.0002	3.59	
2	0.0001	3.54	
2	0.0004	3.48	
2	0.0002	3.48	
2	0.0004	3.42	
2	0.0003	3.27	
2	0.001	3.21	
2	0.0005	3.09	
2	0.0006	3.08	
2	0.0003	3.07	
2	0.0002	3	
2	0.0004	2.97	
2	0.0001	2.94	
2	0.0006	2.89	
2	0.0002	2.87	
2	0.0004	2.86	
2	0.0003	2.81	
2	0.0002	2.8	
2	0.0002	2.76	
3	0.0003	2.56	
3	0.0002	2.45	
2	0.0003	2.41	
3	0.0002	2.29	
3	0.0008	2.25	
3	0.0002	2.15	
4	0.0001	2.13	
3	0.0015	2.13	
3	0.0002	2.12	

3	0.0002	2.08
3	0.0002	1.99
4	0.0002	1.95
4	0.0002	1.95
3	0.0001	1.77
4	0.0001	1.69
4	0.0001	1.66
4	0.0002	1.63
4	0.0001	1.63
4	0.0001	1.53
4	0.0001	1.44
5	0.0001	1.43
9	0.0162	1.37
5	0.0003	1.34

Avg	Avg Coverage	Avg K-mers/Reads	Also found in
1	0.005	37.87	DS253, ND035
1	0.0046	37.51	DS253
1	0.003	26.34	
1	0.0257	25.27	ND035
1	0.0012	24.28	ND035
1	0.0011	23.93	DS253
1	0.0159	23.09	
1	0.0018	18.16	ND035
1	0.0008	17.05	
1	0.0054	8.89	
1	0.0015	8.61	DS253, ND035
1	0.0003	7.34	DS253, ND035
1	0.0003	7.05	
1	0.001	5.96	
1	0.0006	5.58	
1	0.0004	4.98	
1	0.0006	4.84	
2	0.0001	3.33	

Table S4 - Summary results from comparing different sizes of Runs of Homozygosity (ROH). Significa

ROH length threshold	**> 0.1Mb**	**>0.5Mb**	**>1Mb**	**>2Mb**
DS253-ND035-ND036 - DS253	0.005	0.295	0.758	0.666
Nr. of ROHs	5,031	523	109	14
Mean length (Mb)	0.27	0.85	1.6	3.46
Median length (Mb)	0.19	0.68	1.33	2.65
ND035				
Nr. of ROHs	4,836	548	116	13
Mean length (Mb)	0.28	0.84	1.48	2.91
Median length (Mb)	0.2	0.69	1.31	2.61
ND036				
Nr. of ROHs	4,705	559	118	12
Mean length (Mb)	0.28	0.84	1.46	2.75
Median length (Mb)	0.19	0.71	1.29	3.9

Table S5: Genetic load with and without transitions removed

	Sample	HIGH_HOM	HIGH_HET	MODERATE_HOM	MODERATE_HE
Transitions removed	*DS253*	793	100	30245	2572
	ND035	788	101	30213	2424
	ND036	796	100	30164	2539
Transitions kept	*DS253*	2125	307	77448	6445
	ND035	2114	287	77473	6181
	ND036	2118	359	77284	6672

LOW_HOM	LOW_HET	MODIFIER_HOM	MODIFIER_HOM	SNPS_IMPACT_HOM	SNPS_IMPACT_HE
24679	1618	6428085	422839	6483802	427129
24706	1563	6426864	428373	6482571	432461
24634	1641	6422454	428786	6478048	433066
125538	7607	17980551	1242239	18185662	1256598
125583	7575	17977007	1265338	18182177	1279381
125355	7938	17948633	1302163	18153390	1317132

			Genetic load per 100,000 SNPs			
SNPS_ALL_HOM	**SNPS_ALL_HET**	**SNPS_TOTAL**	**HIGH**	**MODERATE**	**LOW**	**MODIFIER**
6483802	427129	6910931	24	912	738	192145
6482571	432461	6915032	24	909	737	192076
6478048	433066	6911114	24	910	737	192063
18185662	1256598	19442260	23	830	1331	191353
18182177	1279381	19461558	23	828	1329	191245
18153390	1317132	19470522	24	828	1328	191055