Pre-extinction Demographic Stability and Genomic Signatures of Adaptation in the Woolly Rhinoceros

Lord, Edana; Dussex, Nicolas; Kierczak, Marcin; Diez-del-Molino, David; Ryder, Oliver A.; Stanton, David W. G.; Gilbert, M. Thomas P.; Sánchez-Barreiro, Fátima; Zhang, Guojie; Sinding, Mikkel Holger S.; Lorenzen, Eline D.; Willerslev, Eske; Protopopov, Albert; Shidlovskiy, Fedor; Fedorov, Sergey; Bocherens, Hervé; Nathan, Senthilvel K. S. S.; Goossens, Benoit; van der Plicht, Johannes; Chan, Yvonne L.; Prost, Stefan; Potapova, Olga; Kirillova, Irina; Lister, Adrian M.; Heintzman, Peter D.; Kapp, Joshua D.; Shapiro, Beth; Vartanyan, Sergey; Götherström, Anders; Dalén, Love

Published in: Current Biology

DOI: 10.1016/j.cub.2020.07.046

Publication date: 2020

Document version
Publisher's PDF, also known as Version of record

Document license: CC BY

Citation for published version (APA):
Pre-extinction Demographic Stability and Genomic Signatures of Adaptation in the Woolly Rhinoceros

Highlights

- Complete genome and mitogenome analysis of the extinct woolly rhinoceros
- Demographic analysis suggests stable population size until close to extinction
- No increased inbreeding or reduced genomic diversity coinciding with human arrival
- Woolly rhinoceros had genetic adaptations to arctic climate

Authors

Edana Lord, Nicolas Dussex, Marcin Kierczak, ..., Sergey Vartanyan, Anders Götherström, Love Dalén

Correspondence

edana.lord@zoologi.su.se (E.L.), love.dalen@nrm.se (L.D.)

In Brief

Here, Lord et al. sequence a complete nuclear genome and 14 mitogenomes from the extinct woolly rhinoceros. Demographic analyses show that the woolly rhinoceros population size was large until close to extinction and not affected by modern human arrival in northeastern Siberia. The extinction may have been mostly driven by climate warming.
SUMMARY
Ancient DNA has significantly improved our understanding of the evolution and population history of extinct megafauna. However, few studies have used complete ancient genomes to examine species responses to climate change prior to extinction. The woolly rhinoceros (*Coelodonta antiquitatis*) was a cold-adapted megaherbivore widely distributed across northern Eurasia during the Late Pleistocene and became extinct approximately 14 thousand years before present (ka BP). While humans and climate change have been proposed as potential causes of extinction [1–3], knowledge is limited on how the woolly rhinoceros was impacted by human arrival and climatic fluctuations [2]. Here, we use one complete nuclear genome and 14 mitogenomes to investigate the demographic history of woolly rhinoceros leading up to its extinction. Unlike other northern megafauna, the effective population size of woolly rhinoceros likely increased at 29.7 ka BP and subsequently remained stable until close to the species’ extinction. Analysis of the nuclear genome from a 18.5-ka-old specimen did not indicate any increased inbreeding or reduced genetic diversity, suggesting that the population size remained steady for more than 13 ka following the arrival of humans [4]. The population contraction leading to extinction of the woolly rhinoceros may have thus been sudden and mostly driven by rapid warming in the Bølling-Allerød interstadial. Furthermore, we identify woolly rhinoceros-specific adaptations to arctic climate, similar to those of the woolly mammoth. This study highlights how species respond differently to climatic fluctuations and further illustrates the potential of palaeogenomics to study the evolutionary history of extinct species.

RESULTS AND DISCUSSION
Genome Sequencing
To investigate changes in genetic diversity that preceded the extinction of the woolly rhinoceros, and to obtain a glimpse of the species’ genomic adaptation to the arctic environment, we sequenced a woolly rhinoceros nuclear genome and 14 mitochondrial genomes that ranged in age from >50 to 14.1 thousand calibrated years before present (ka cal BP). The specimen from which we recovered the nuclear genome was...
radiocarbon dated to 18,530 ± 170 cal BP (Data S1A) and had an endogenous DNA content of 56.7%. After mapping the raw data to a new and high-quality assembly of its closest extant relative, the Sumatran rhinoceros (*Dicerorhinus sumatrensis*) [5, 6], the average genome coverage was 13.8X, with 70% of the genome having a coverage ≥ 10X. The average DNA fragment length was 84 bp, and overall, we identified 28,180,718 high-quality SNPs after filtering out SNPs with low mapping and base quality and low coverage (see STAR Methods). Furthermore, we conducted low-coverage shotgun sequencing on 13 additional woolly rhinoceros specimens, recovering in total 14 full mitochondrial genomes from northeastern Siberia (Figure 1A) with an average depth ranging from 7.5X to 912.8X (Data S1A).

**Demographic History**

Among the 14 mitochondrial genomes, we identified 13 unique mitogenome haplotypes (Figure S1A). Based on both a Bayesian phylogeny and median-joining network, we identified two clades (1 and 2) that diverged ~205 ka BP (95% highest posterior density [HPD]: 440–116 ka BP) and persisted up until close to the extinction of the woolly rhinoceroses (Figures 1B and S1A). There was no indication of geographic or temporal structuring between or within these clades, except for a single sample from Wrangel Island forming its own distinct lineage, diverging from clade 1 ~154 ka BP (95% HPD: 326–91 ka BP). This unique lineage may result from isolation on Wrangel Island because of less favorable habitat south of this locality, as steppe into forested and shrub lowlands during warm periods [9, 10]. Future studies that include additional samples from Wrangel will allow exploration of potential genetic structure between Wrangel and adjacent regions in northeastern Siberia.

The mitochondrial phylogeny had long, well-resolved branches between the three clades (Figure 1B), similar to the pattern observed in woolly mammoths (*Mammuthus primigenius*) [11]. It has been hypothesized that this mitogenome structure in mammoths potentially resulted from isolation in interglacial refugia [11]. Finding a similar structure in woolly rhinoceroses and mammoths suggests that these species responded similarly to past climate warming. Within clades 1 and 2 of the mitochondrial phylogeny, the short unresolved branches suggest a recent diversification approximately 86–22 ka BP (95% HPDs for Node C and D, Figure 1B). In the demographic analyses, a model of constant population size obtained the highest support (Figure S1B; Data S1A), with a female effective population size (*N*<sub>ef</sub>) of around 100,000 over the last 110 ka until the extinction at ca. 14 ka BP. However, an alternative model, albeit with a lower likelihood of support, indicated an expansion in *N*<sub>ef</sub> consistent with the recent diversification of lineages within each clade observed in the phylogeny (Figure S1C).

To further examine the demographic history of the woolly rhinoceroses, we used a pairwise sequentially Markovian coalescent (PSMC) analysis based on the nuclear genome (Figure 2). Effective population size (*N*<sub>e</sub>) increased gradually from ~1 Ma BP during the Early Pleistocene, reaching a peak of ~21,000 at around 152 ka BP (95% confidence interval [CI]: 274–111 ka BP), during the Marine Isotope Stage 6 (MIS 6) glaciation (130–191 ka BP). Subsequently, *N*<sub>e</sub> decreased 10-fold from 127 ka BP (95% CI: 226–94 ka BP) until 29.7 ka BP (95% CI: 40–26.3 ka BP), at which point there was a rapid expansion in *N*<sub>e</sub>. The effective population size then remained constant until the time of the death of the individual (18.5 ka cal BP), approximately 4.5 ka prior to the extinction of the species.
The observation that \( N_{eW} \) was higher than \( N_e \) during the Late Pleistocene (Figures 2 and S1B) could potentially be explained by male-biased dispersal and female philopatry. However, there is little evidence for sex-biased dispersal in extant rhinoceroses (e.g., black rhinoceros \( Diceror bicornis \) [14]; white rhinoceros \( Ceratotherium simum \) [15]), which makes this explanation unlikely for woolly rhinoceros. Instead, we hypothesize that the comparatively high \( N_{eW} \) in the woolly rhinoceros is a consequence of a high variance in male reproductive success, similar to what has been reported for white rhinoceros [16]. Future analysis of nuclear data from multiple male and female woolly rhinoceros will be necessary to further explore this question and to provide genomic insights into aspects of its behavior.

The observed increase in \( N_e \) leading up to MIS 6 may signify a demographic expansion but could alternatively be attributed to population subdivision and the divergence of the two clades identified in the mitochondrial analyses. It is plausible that these clades formed in allopatry, possibly during an interglacial period, and that these populations subsequently expanded and merged during or after MIS 6, leading to the lack of phylogeographic structure observed in the mitochondrial data. Thus, the peak observed at MIS 6 may be an artifact of a population subdivision rather than an increase in population size as population structure is known to affect PSMC [17–19]. Following MIS 6, the effective population size decreased through the Eemian interglacial (~130–115 ka) and the beginning of the last glacial period, reaching a minimum \( N_e \) at ~33 ka BP.

Although PSMC has reduced power in estimating \( N_e \) during the 20 ka prior to the age of the sample [17], we observed an increase in \( N_e \) at 29.7 ka BP. While this increase is consistent with previous estimates based on short mitochondrial DNA sequences from woolly rhinoceroses [2] and the diversification within each clade observed in our mitogenomes, it is in contrast to data from the woolly mammoth, which did not indicate an expansion in \( N_e \) at that time [20]. We hypothesize that the observed increase in \( N_e \) in the woolly rhinoceros may have been related to the transition from the climatically unstable MIS 3 to the more stable cold period of MIS 2 around 29 ka BP [21], which was a period suggested to have suitable habitat in northeastern Siberia for cold-adapted species [22, 23]. However, as the woolly rhinoceros experienced an increase in \( N_e \), other cold-adapted taxa such as the woolly mammoth’s \( N_e \) remained stable. MIS 2 may thus have provided a particularly suitable habitat for the woolly rhinoceros with glacial tundra-steppe conditions prevailing and allowing population expansion [24]. An alternative explanation may be that the increase in \( N_e \) represents the merging of populations as the range of the highly specialized grazers such as woolly rhinoceroses [24] contracted toward northeastern Siberia [18], while the mammoth, which may have been ecologically more flexible as exemplified by its wider distribution [2, 25, 26], maintained a constant \( N_e \).

Although tentative because of the limitations of PSMC analysis, our results suggest that the woolly rhinoceros’ population size may have remained constant after the expansion 29.7 ka BP and until the death of the sequenced individual. Our mitochondrial data further supports a scenario of population stability until close to the extinction of the species (Figures 1B and S1B), since the two lineages identified here persisted until within ~300 years of the estimated extinction event at ~14 ka BP [1]. In spite of a progressive range contraction toward northeastern Siberia, the woolly rhinoceros remained stable in population size until very recently.
starting 35 ka BP, the fossil record indicates that the species was still widespread up until 18.5 ka BP [1], which may explain why our population size estimates remained constant. Interestingly, data from several other mammals highlight the importance of northeastern Siberia as a late glacial refugium. For example, recent analyses indicate that extant wolf lineages originated in northeastern Siberia [27], and it has been hypothesized that admixing of modern human populations occurred in the area, prior to the colonization of North America (e.g., [4, 28, 29]). Similarly, horse, bison, and collared lemming also have highly diverged mitochondrial lineages that survived in northeastern Siberia after the Last Glacial Maximum [30–32], suggesting long-term population continuity for several taxa in this region.

Genomic Diversity and Extinction

The woolly rhinoceros genome had an average heterozygosity of approximately 1.7 heterozygous sites per 1,000 bp (95% CI: 1.66–1.74). This is higher than the genomic diversity observed in a previously published mainland mammoth genome (1.25 heterozygous sites per 1,000 bp [20]), as well as the extant Sumatran rhinoceros (1.3 heterozygous sites per 1,000 bp [12]) and Northern and Southern White rhinoceros (1.1 and 0.9 heterozygous sites per 1,000 bp, respectively [33]). Based on identification of runs of homozygosity (ROH), we estimated the inbreeding coefficient (F_{ROH}) to be 5.9% when considering ROH regions as defined by USER treatment (see STAR Methods; Figure S2). This observed level of inbreeding is comparatively low and, for example, on par with non-African human populations [34]. We note, however, that the level of inbreeding is higher than that observed in a Late Pleistocene mainland mammoth (F_{ROH} 0.83%) and indicates some degree of background relatedness from mating between distant relatives [35], potentially because of higher population substructure and/or reduced local population size at the time compared with the mammoth. However, this result is in stark contrast to a 4.3 ka BP mammoth from Wrangel Island (F_{ROH} 23.3%), which showed increased inbreeding associated with long-term small population size [20, 36].

Taken together, our analyses of nuclear and mitochondrial genomic diversity in the woolly rhinoceros provide no evidence for a decline in population size preceding the samples analyzed here, nor any indication of elevated inbreeding typical of small populations. While we cannot exclude the role of humans in woolly rhinoceros’ extinction, our results imply that the arrival of anatomically modern humans in northeastern Siberia was not correlated with a demographic decline in the woolly rhinoceros. However, we caution that the earliest evidence of human presence in northeastern Siberia, dated to 31.6 ka BP [4], may represent temporary settlements [37] and that currently there is only evidence of sporadic human presence through MIS 3–2 [38], thus humans may have only had a limited negative impact on woolly rhinoceros populations.

Overall, our findings of a stable population size until at least 18.5 ka suggest that the final decline toward extinction was rapid and started within the 4,500 years prior to the extinction (i.e., after the death of the individual whose genome was sequenced here). This severe and rapid demographic decline, which based on radiocarbon evidence [1] likely coincided with the Bolling-Allerød interstadial (14.6–12.8 ka), could imply that the extinction of woolly rhinoceros was primarily driven by the changes in climate and vegetation characteristic of the period [22]. Across Eurasia, the Bolling-Allerød interstadial was characterized by an increase in forest habitats and woody plant cover [9], Stuart and Lister [1] previously suggested that the replacement of low-growing vegetation by shrub-tundra and tree biomes (e.g., Salix sp., Betula sp.) in Siberia during the warm Bolling-Allerød interstadial [9, 20], combined with increased snowfall [39], likely led to the extinction of the woolly rhinoceros.
extinction of the woolly rhinoceros. Additional sequencing of individuals closer to the extinction event will be needed to gain a better understanding of the timing and rate of decline toward extinction.

Adaptation to Cold Environments

We undertook a preliminary evaluation of adaptations in woolly rhinoceros relative to Sumatran rhinoceros by examining non-synonymous mutations (i.e., missense; loss of function, LoF) across 19,556 coding genes. Overall, we found 1,524 identifiable genes with non-synonymous mutations (n missense = 1,386, n LoF = 138; Data S1D) associated with biological processes including cellular component organization or biogenesis, cellular process, localization, reproduction, biological regulation, response to stimulus, developmental processes, and metabolic processes, several of which are significantly overrepresented (Data S1E and S1F). In contrast to previous analyses of another cold-adapted megafaunal species, the woolly mammoth [40], we did not observe non-synonymous variants in genes associated with fat deposition and changes to circadian rhythm thought to have played a role in mammoth adaptation to the arctic environment. However, in 89 genes, the woolly mammoth and the woolly rhinoceros both had non-synonymous variants potentially indicative of positive selection, including in TRPA1 (Transient Receptor Potential subfamily A; Data S1G and S1H), which is known to be involved in adaptation to cold tolerance [41, 42]. Variants undergoing positive selection in genes encoding TRP channels, including TRPA1, have recently been described in a range of cold-adapted taxa [40, 43]. Furthermore, there was one gene (KCNK17, potassium channel subfamily K) in which both species have a LoF mutation. KCNK17 is a paralog of KCNK4 (also known as TRAAK, TWIK-Related Arachidonic Acid-Stimulated Potassium Channel Protein), which under normal function has been shown to silence TRP proteins including TRPA1 and TRPM8 [42]. Thus, this gene is involved in cold temperature perception and, when knocked out, may play a role in cold adaptation [42].

To further identify genes that may have been of adaptive significance in the woolly rhinoceros, we ranked all identifiable missense mutations (n = 17,888) according to three indices (amino acid index [aaI], experimental exchangeability, and Sneath’s amino acid dissimilarity) in order to evaluate their impact on protein structure and physicochemical properties (Data S1C; see STAR Methods). The results showed that all three indices gave similar results (Data S1C). The distribution of aaI was bi-modal, with the majority of mutations predicted to cause mostly weak to moderate changes in protein structure (Figure S3). However, there were 284 variants with an aaI of 1, indicating maximal change of amino acid physicochemical properties. Of these variants, 83 were found across 41 different olfactory receptor genes (Data S1C), which is consistent with frequent gene gains and losses during the evolution of this gene family [44].

Conclusions

Our analyses of genomic diversity have several implications for understanding the population history and biology of the woolly rhinoceros. First, the finding of deep divergence among mitochondrial lineages hints at a dynamic history during the Middle Pleistocene, possibly characterized by the fragmentation and subsequent merger of populations. Second, analyses of mitogenomes and the nuclear genome both suggest that the species’ final decline toward extinction was rapid and did not begin until after 18.5 ka BP. This implies that the woolly rhinoceros Ne did not start to decline until approximately 13 ka after the first arrival of humans in northeastern Siberia [4, 28, 29]. This does not exclude the possibility that humans later contributed to their extinction. For instance, hunting of woolly rhinoceros by humans could have reduced the population growth rate and, thus, may have accelerated the extinction of the species. However, given the data at hand, it appears likely that changes in the
environment, associated with the onset of the Bolling-Allerød interstadial, were the primary drivers of the woolly rhinoceros’ extinction. It should be possible to further investigate the extent to which the final demographic decline coincided with the Bolling-Allerød by analyzing additional genomes from the time period 18–14 ka BP. Finally, our preliminary assessment of adaptive genetic variation in the woolly rhinoceros identified a range of non-synonymous changes in genes associated with several biological processes, including a gene (TRPA1) involved in temperature sensation. Taken together, these findings highlight the utility of genomic data in unraveling previously unknown evolutionary processes in extinct species and illustrate the need to investigate demographic trajectories in other megafauna to develop a better understanding of the timing and rate of demographic change during the Late Quaternary.

STAR METHODS

Detailed methods are provided in the online version of this paper and include the following:

- **KEY RESOURCES TABLE**
- **RESOURCE AVAILABILITY**
  - Lead Contact
  - Materials Availability
  - Data and Code Availability
- **EXPERIMENTAL MODEL AND SUBJECT DETAILS**
- **METHOD DETAILS**
  - DNA extraction
  - Library preparation
- **QUANTIFICATION AND STATISTICAL ANALYSIS**
  - De novo reference assembly and annotation
  - Estimation of endogenous DNA content
  - Mitogenome data processing
  - Nuclear genome data processing
  - Mitogenome data analysis
  - Demographic reconstruction
  - Heterozygosites and inbreeding
  - Non-synonymous mutations

SUPPLEMENTAL INFORMATION

Supplemental Information can be found online at https://doi.org/10.1016/j.cub.2020.07.046.

ACKNOWLEDGMENTS

We thank Anthony Stuart for providing fossil samples. The authors are grateful to Alexander Banderov, Semen Ivanov, Klimovskiy Aisen, and several others for help with collection of samples in the field. We acknowledge support from the Uppsala Multidisciplinary Centre for Advanced Computational Science for assistance with massively parallel sequencing and access to the UPPMAX computational infrastructure. We also acknowledge Tom van der Valk for sharing a script for CpG sites filtering and Shyam Gopalakrishnan for sharing a demultiplex script of BGI raw data. Sequencing was performed by the Swedish National Genomics Infrastructure (NGI) at the Science for Life Laboratory (ILLUMINA) which is supported by the Swedish Research Council (Vetenskapsrådet). We thank the Knut and Alice Wallenberg Foundation for sharing a demultiplex script of BGI raw data. Sequencing was performed by the Swedish National Genomics Infrastructure (NGI) at the Science for Life Laboratory (ILLUMINA), and the Knut and Alice Wallenberg Foundation, and at the China National Genomics Centre (BGISeq data). We thank the Sabah Biodiversity Centre for allowing us to export samples from Sumatran rhinoceros individuals from Sabah (License Ref JKM/MBS.1000-2/3 JLD.2 (32)), and also thank Dr. Zainal Zainuddin and Dr. John Payne from Borneo Rhino Alliance (BORA) and the late Dr. Diana Angeles Ramirez Saldivar from the Wildlife Rescue Unit for providing samples from the Sumatran rhinoceros individual used for the reference genome assembly. This work was supported by Formas (2015-676 and 2018-01640) and the Bolin Centre for Climate Research to L.D., the Swiss National Science Foundation (P2SKP3_165031 and P300PA_177845) and the Carl Tryggers Foundation (CTS 19:257 to N.D., the Carl Tryggers Foundation (CTS 17:109) to D.D.-d.-M., and an ERC Consolidator Award 681396-Extinction Genomics to M.T.P.G. M.K. is financially supported by the Knut and Alice Wallenberg Foundation as part of the National Bioinformatics Infrastructure Sweden at SciLifeLab.

AUTHOR CONTRIBUTIONS


DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: May 14, 2020
Revised: June 18, 2020
Accepted: July 14, 2020
Published: August 13, 2020

REFERENCES


# STAR METHODS

## KEY RESOURCES TABLE

<table>
<thead>
<tr>
<th>REAGENT or RESOURCE</th>
<th>SOURCE</th>
<th>IDENTIFIER</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Chemicals, Peptides, and Recombinant Proteins</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EDTA</td>
<td>ThermoFisher Scientific</td>
<td>Cat#15575020</td>
</tr>
<tr>
<td>UREA</td>
<td>VWR</td>
<td>Cat#443874G</td>
</tr>
<tr>
<td>DTT</td>
<td>ThermoFisher Scientific</td>
<td>Cat#R0861</td>
</tr>
<tr>
<td>Tango Buffer (10X)</td>
<td>ThermoFisher Scientific</td>
<td>Cat#BY5</td>
</tr>
<tr>
<td>ATP (100mM)</td>
<td>ThermoFisher Scientific</td>
<td>Cat#R0441</td>
</tr>
<tr>
<td>USER Enzyme</td>
<td>NEB</td>
<td>Cat#M5505L</td>
</tr>
<tr>
<td>T4 DNA Ligase (5U/ul)</td>
<td>ThermoFisher Scientific</td>
<td>Cat#EL0011</td>
</tr>
<tr>
<td>Bst Polymerase</td>
<td>NEB</td>
<td>Cat#M0275S</td>
</tr>
<tr>
<td>AccuPrime Pfx</td>
<td>Life Technologies</td>
<td>Cat#12344-024</td>
</tr>
<tr>
<td>T4 DNA ligase (400U/ul)</td>
<td>NEB</td>
<td>Cat#M0202S</td>
</tr>
<tr>
<td>T4 DNA polymerase (3U/ul)</td>
<td>NEB</td>
<td>Cat#M0203S</td>
</tr>
<tr>
<td>Bst 2.0 Warmstart Polymerase (8U/ul)</td>
<td>NEB</td>
<td>Cat#M0538S</td>
</tr>
<tr>
<td>NaCl 5M solution</td>
<td>Sigma-Aldrich</td>
<td>Cat#59222C-500ML</td>
</tr>
<tr>
<td>PEG-8000</td>
<td>Sigma-Aldrich</td>
<td>Cat#89510-250G-F</td>
</tr>
<tr>
<td>BSA (20 mg/mL)</td>
<td>NEB</td>
<td>Cat#B9000S</td>
</tr>
<tr>
<td><strong>Critical Commercial Assays</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>High Sensitivity DNA kit</td>
<td>Agilent</td>
<td>Cat#5067-4626</td>
</tr>
<tr>
<td>KingFisher™ Cell and Tissue DNA Kit</td>
<td>ThermoFisher Scientific</td>
<td>Cat#97030196</td>
</tr>
<tr>
<td><strong>Deposited Data</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Raw fastq reads (mitochondrial and nuclear data)</td>
<td>This study</td>
<td>ENA study accession number PRJEB35556</td>
</tr>
<tr>
<td>de-novo assembly for Dicerorhinus sumatrensis</td>
<td>This study</td>
<td>GenBank: JABWHU000000000</td>
</tr>
<tr>
<td><strong>Oligonucleotides</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IS1 adaptor P5:</td>
<td>[45]</td>
<td>Sigma-Aldrich</td>
</tr>
<tr>
<td>5'-&gt;3'</td>
<td>A<em>C</em>A<em>C</em>TCTTTCCCTACACGACGCTCTTC</td>
<td></td>
</tr>
<tr>
<td>IS2 adaptor P7:</td>
<td>[45]</td>
<td>Sigma-Aldrich</td>
</tr>
<tr>
<td>5'-&gt;3'</td>
<td>G<em>T</em>G<em>A</em>CTGGAGTTCAGACGTGTGCTCTTC</td>
<td></td>
</tr>
<tr>
<td>IS3 adaptor P5+P7:</td>
<td>[45]</td>
<td>Sigma-Aldrich</td>
</tr>
<tr>
<td>5'-&gt;3'</td>
<td>A<em>G</em>A<em>T</em>CGGAA<em>G</em>A<em>G</em>C</td>
<td></td>
</tr>
<tr>
<td>Illumina AmplifyingPrimer IS4:</td>
<td>[45]</td>
<td>Sigma-Aldrich</td>
</tr>
<tr>
<td>5'-&gt;3'</td>
<td>AATGATACGGCGACCACCGAGATCTACACTC</td>
<td></td>
</tr>
<tr>
<td>Illumina Indexing Primer:</td>
<td>[45]</td>
<td>Sigma-Aldrich</td>
</tr>
<tr>
<td>5'-&gt;3'</td>
<td>CAAGCAGAAGACGGCATACGAGATNNNNN</td>
<td></td>
</tr>
<tr>
<td>Ns represent indexes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>IS3 ATDC3 adaptor:</td>
<td>[46]</td>
<td>Sigma-Aldrich</td>
</tr>
<tr>
<td>5'-&gt;3'</td>
<td>G<em>AT</em>C<em>GGAA</em>G<em>A</em>G*C[C3spacer]</td>
<td></td>
</tr>
</tbody>
</table>

(Continued on next page)
**REAGENT or RESOURCE** | **SOURCE** | **IDENTIFIER**
--- | --- | ---
BGISEQ adaptor AD1_Long: 5’→3’ AAGCAGAAGACGGCATACGAGATGTT GTCTTCTAAGACGGCATACGAGATGTT | [47] Sigma-Aldrich |
BGISEQ adaptor AD1_Short: 5’→3’ AAGTCGGAGGCC | [47] Sigma-Aldrich |
BGISEQ adaptor AD2_Long: 5’→3’ TTGTCTTCTTAAGGAACGACATGGCT ACGATCCGACTT | [46] Sigma-Aldrich |
BGISEQ adaptor AD2_Short: 5’→3’ AAGTCGGATCGT | [46] Sigma-Aldrich |
BGISEQ Indexing primer: 5’→3’ TGTGAGCCAAGGAGTTGNNNNNNN NNTTGTCTTCTAAGACCNG | [46] Sigma-Aldrich |
Common amplifying primer BGI forward: 5’→3’ GAACGACATGGCTACGA | [46] Sigma-Aldrich |
Software and Algorithms

<table>
<thead>
<tr>
<th>Tool</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>OxCal v4.3</td>
<td>[48] <a href="https://c14.arch.ox.ac.uk/oxcal.html">https://c14.arch.ox.ac.uk/oxcal.html</a></td>
</tr>
<tr>
<td>HiRise pipeline</td>
<td>[50] Dovetail Genomics</td>
</tr>
<tr>
<td>BLAST+ 2.5.0</td>
<td>[51] NCBI</td>
</tr>
<tr>
<td>MESPA pipeline</td>
<td>[52] <a href="https://sourceforge.net/projects/mespa/">https://sourceforge.net/projects/mespa/</a></td>
</tr>
<tr>
<td>cufflinks v 2.2.1</td>
<td>[53] <a href="http://cole-trapnell-lab.github.io/cufflinks/">http://cole-trapnell-lab.github.io/cufflinks/</a></td>
</tr>
<tr>
<td>eggNOG-mapper v4.5.1</td>
<td>[54] <a href="http://eggnog-mapper.embl.de/">http://eggnog-mapper.embl.de/</a></td>
</tr>
<tr>
<td>bcfl2fastq v1.8.3</td>
<td>Illumina <a href="https://support.illumina.com/sequencing/sequencing_software/bcl2fastq-conversion-software.html">https://support.illumina.com/sequencing/sequencing_software/bcl2fastq-conversion-software.html</a></td>
</tr>
<tr>
<td>Custom BGI demultiplexing script</td>
<td>Shyam Gopalakrishnan <a href="https://github.com/shyamsg/SantasHelpers/">https://github.com/shyamsg/SantasHelpers/</a></td>
</tr>
<tr>
<td>SeqPrep</td>
<td>John St. John <a href="https://github.com/jstjohn/SeqPrep">https://github.com/jstjohn/SeqPrep</a></td>
</tr>
<tr>
<td>BWA v0.7.13</td>
<td>[55] <a href="http://bio-bwa.sourceforge.net/">http://bio-bwa.sourceforge.net/</a></td>
</tr>
<tr>
<td>SAStools v1.3</td>
<td>[56] <a href="https://sourceforge.net/projects/samtools/files/samtools/1.3/">https://sourceforge.net/projects/samtools/files/samtools/1.3/</a></td>
</tr>
<tr>
<td>Geneious® v7.0.336</td>
<td>[57] <a href="https://www.geneious.com/">https://www.geneious.com/</a></td>
</tr>
<tr>
<td>Picard v1.141</td>
<td>Broad Institute <a href="http://broadinstitute.github.io/picard">http://broadinstitute.github.io/picard</a></td>
</tr>
<tr>
<td>GATK v3.4.0</td>
<td>[58] <a href="https://gatk.broadinstitute.org/hc/en-us">https://gatk.broadinstitute.org/hc/en-us</a></td>
</tr>
<tr>
<td>Map damage v2.0</td>
<td>[59] <a href="https://ginolhac.github.io/mapDamage/">https://ginolhac.github.io/mapDamage/</a></td>
</tr>
<tr>
<td>Qualimap v2.2.1</td>
<td>[60] <a href="http://qualimap.bioinfo.cipf.es/">http://qualimap.bioinfo.cipf.es/</a></td>
</tr>
<tr>
<td>BEDtools v2.29.2</td>
<td>[61] <a href="https://bedtools.readthedocs.io/en/latest/">https://bedtools.readthedocs.io/en/latest/</a></td>
</tr>
<tr>
<td>jModelTest v2.1.9</td>
<td>[64] <a href="https://github.com/ddarriba/jmodeltest2">https://github.com/ddarriba/jmodeltest2</a></td>
</tr>
<tr>
<td>Figtree v1.4.4</td>
<td>[65] <a href="http://tree.bio.ed.ac.uk/software/figtree/">http://tree.bio.ed.ac.uk/software/figtree/</a></td>
</tr>
<tr>
<td>Tracer v1.7.1</td>
<td>[66] <a href="https://github.com/beast-dev/tracer/releases/tag/v1.7.1">https://github.com/beast-dev/tracer/releases/tag/v1.7.1</a></td>
</tr>
<tr>
<td>PSMC v0.6.5</td>
<td>[17] <a href="https://github.com/lh3/psmc">https://github.com/lh3/psmc</a></td>
</tr>
<tr>
<td>mlRho v2.7</td>
<td>[67] <a href="http://guanine.evolbio.mpg.de/mlRho/">http://guanine.evolbio.mpg.de/mlRho/</a></td>
</tr>
<tr>
<td>PLINK v1.9</td>
<td>[68] <a href="http://www.cog-genomics.org/plink2">http://www.cog-genomics.org/plink2</a></td>
</tr>
</tbody>
</table>

(Continued on next page)
RESOURCE AVAILABILITY

Lead Contact
Further information and requests for reagents and data may be directed to and will be fulfilled by the Lead Contact, Edana Lord (edana.lord@zoologi.su.se).

Materials Availability
This study did not generate new unique reagents.

Data and Code Availability
Raw fastq reads for mitogenome and nuclear data are deposited at the European Nucleotide Archive (ENA; study accession number PRJEB35556). The de-novo assembly for Dicerorhinus sumatrensis is deposited on GenBank (accession: JABWHU00000000).

EXPERIMENTAL MODEL AND SUBJECT DETAILS

We obtained 12 bones, one mummified tissue biopsy, and one hair sample of woolly rhinoceros, which were radiocarbon dated to between 14,100 and > 50,000 cal BP from North-eastern Siberia (Data S1A). Radiocarbon dating was performed at the Oxford Radiocarbon Accelerator Unit (ORAU, OxA), Beta Analytics (Miami, FL), ETH Zürich, and the Center for Isotope Research of Groningen University (GrA). We calibrated all radiocarbon dates using the IntCal13 calibration curve [48] in OxCal v4.3 [71]. For the de-novo assembly, we obtained tissue and cell lines from one male Sumatran Rhinoceros (Dicerorhinus sumatrensis), called Kertam, that originated from Borneo.

METHOD DETAILS

DNA extraction
We extracted DNA from bone samples according to protocol C in Yang et al. [72] as modified in Brace et al. [73]. For the mummified tissue biopsy and hair samples, we extracted DNA following Dabney et al. [74], but substituted the digestion buffer and incubation temperature with that described in Gilbert et al. [75]. Appropriate precautions were taken to minimize the risk of contamination during the processing of ancient samples [76].

For the de-novo assembly, we extracted DNA from blood and cell lines from Kertam, using a Kingfisher robot (Thermo Fisher Scientific) and following the Kingfisher blood & tissue extraction protocol according to the manufacturer’s instructions. Concentrations were measured using QuBit®/C210 Fluorometer (Invitrogen, USA) and the quality of the DNA was evaluated by running the samples through agarose gels with electrophoresis.

Library preparation
Double stranded Illumina libraries were built for 14 extracts according to Meyer & Kircher [45], along with 2 extraction blanks. 20 μl of DNA extract was used in a 40 μl blunt-end repair reaction with the following final concentration: 1 × buffer Tango, 100 μM of each dNTP, 1 nM ATP, 25 U T4 polynucleotide kinase (Thermo Scientific) and 3U USER enzyme (New England Biolabs). A USER enzyme treatment was performed to excise uracil residues resulting from post-mortem damage [77, 78]. Samples were incubated for 3 h at 37°C, followed by the addition of 1 μl T4 DNA polymerase (Thermo Scientific) and incubation at 25°C for 15 min and 12°C for 5 min. The samples were then purified using MinElute spin columns following the manufacturer’s protocol and eluted in 20μl EB buffer. Next, an adaptor ligation step was performed where DNA fragments within each library were ligated to a combination of incomplete, partially double-stranded P5- and P7-adapters (10 μM each). This reaction was performed in a 40 μl reaction volume using 20 μl of blunt-ended DNA library and 1 μl P5-P7 adaptor mix per sample with a final concentration of 1 × T4 DNA ligase buffer, 5% PEG-4000, 5U T4 DNA ligase (Thermo Scientific). Samples were incubated for 30 min at room temperature and cleaned using MinElute spin columns as described above.

Continued
Next, we performed an adaptor fill-in reaction in 40 µl final volume using 20 µl adaptor ligated DNA with a final concentration of 1 × Thermopol Reaction Buffer, 250 µM of each dNTP, 8U Bst Polymerase, Long Fragments. The libraries were incubated at 37°C for 20 min, and then heat-inactivated at 80°C for 20 min. These libraries were then used as stock for indexing PCR amplification for screening (i.e., estimation of endogenous DNA content of each sample) and deep-sequencing.

PCR amplifications were performed in 25 µl volumes with 3µl of adaptor-ligated library as template, with the following final concentrations: 1x AccuPrime reaction mix, 0.3µM IS4 amplification primer, 0.3µM P7 indexing primer, 7 U AccuPrime Pfx (Thermo Scientific) and the following cycling protocol: 95°C for 2 min, 12 cycles at 95°C for 30 s, 55°C for 30 s and 72°C for 1 min and a final extension at 72°C for 5 min. We used dual unique indexes of 6 bp for each library.

Purification and size selection of the libraries were performed using Agencourt AMPure XP beads (Beckman Coulter, Brea, CA, USA), first using a 0.5X bead:DNA ratio and second with a 1.8X bead:DNA ratio to remove long and short (i.e., adapter dimers) fragments, respectively. Library concentration was measured with a high-sensitivity DNA chip on a Bioanalyzer 2100 (Agilent, Santa Clara, CA, USA). Finally, multiplexed libraries were pooled into a single pool in equimolar concentrations and sequenced on Illumina HiSeq2500 in High Output mode with a 2 × 125bp setup at SciLifeLab, Stockholm. The nuclear genome sample, ND035, was deep sequenced on an Illumina HiSeqX with a 2 × 150 bp setup at SciLifeLab, Stockholm.

For the hair sample, ND045, a double stranded library was constructed following the Meyer & Kircher protocol [45] as modified by Heintzman et al. [79]. We also built additional libraries for three samples (ND012, ND036, ND045) using the BEST2 library build protocol, a blunt-end, single tube library preparation procedure suitable for degraded DNA samples [46] and using custom-design adaptor oligos specific for the BGISEQ-500 Sequencing Platform [47]. No USER treatment was performed for these three libraries.

**QUANTIFICATION AND STATISTICAL ANALYSIS**

**De-novo reference assembly and annotation**

We generated a de novo reference genome for one male Sumatran rhinoceros (*D. sumatrensis*) by sequencing a combination of Chicago, Hi-C, mate-pair and short insert libraries from a high molecular weight DNA extract from one male from the Bornean population (Kertam). An initial assembly based on the short insert and mate-pairs was done using Allpaths v.2.0 [49]. The final genome assembly was done using the HiRise pipeline (Dovetail Genomics, [50]). The final assembly size was of 2.4 Gb and comprised 1,763 scaffolds with an N50 of 62 Mb, where 99% of the genome was comprised within 44 scaffolds > 1 Mb.

We identified X-linked scaffolds in the Sumatran rhinoceros genome in BLAST+ 2.5.0 [51] using the horse X chromosome as subject sequence. The BLAST+ parameters were set as: -evalue = 1e-10; -word_size = 15; -max_target_seqs = 1000. For all downstream analyses, we excluded two X chromosome-linked scaffolds (Sc9M7eS_1319; HRSCAF = 1962 and Sc9M7eS_931; HRSCAF = 1475) from the assembled genome.

We annotated the assembly using the MESPA pipeline [52]. We collapsed reference protein sets for white rhinoceros (*Ceratotherium simum simum*; GenBank: GCF_000283155.1) to 90% coverage following Uniprot90 guidelines where each protein cluster is composed of sequences with at least 90% sequence identity to, and 80% overlap with, the longest sequence using a custom script. In that way, we discarded isoforms of the reference datasets. We then used MESPA to extract the gene models in Sumatran rhinoceros with 90% length coverage to each set of reference proteins and to generate an annotation in gff format. We extracted 99% (21,953 out of 22,054) high quality protein models (i.e., aligning to 90% of their expected length) using white rhinoceros as a reference protein set.

We extracted the CDSs and protein sequences of this annotation with cufflinks v 2.2.1 [53, 80] gffread command for downstream analyses using the -V option to remove gene models with in-frame STOP codons. We retained 19,556 gene models with a mean length of 1,724 bp (Median = 1,051; min = 34; max = 26,418).

Finally, we performed a functional annotation of these gene models using the eggNOG-mapper v4.5.1 [54]. We used ‘Mammals’ as a taxonomic scope and the ‘Restrict to one-to-one’ and the ‘Use experimental terms only’ to prioritize precision and quality of matches.

**Estimation of endogenous DNA content**

Raw Illumina sequence data were demultiplexed based on their unique indices from Bcl to Fastq using bcl2Fastq v1.8.3 (CASAVA software suite) while raw BGI data was converted and samples were demultiplexed using a custom script (https://github.com/shyamsg/SantasHelpers/). We used SeqPrep (https://github.com/jstjohn/SeqPrep) to trim adapters and merge paired-end reads using default settings, with a minor modification in the source code that allowed us to choose the best quality scores of bases in the merged region instead of aggregating the scores, following [20]. Raw Illumina sequencing reads were aligned to the de-novo genome of the Sumatran rhinoceros (*Dicerorhinus sumatrensis*), which is the closest extant relative to woolly rhinoceros, with BWA v0.7.13 [55] and then processed with SAMTools v1.3 [56]. We mapped the merged sequencing reads against the reference genome using the BWA aln algorithm and slightly modified default settings with deactivated seeding (-l 16,500), allowing more substitutions (-n 0.01) and allowing up to two gaps (-o 2). We then used the BWA samse command to generate alignments and subsequently converted reads mapping to the reference genome from SAM to BAM format, sorted and indexed using SAMTools. We estimated the endogenous DNA content for each sample as the proportion of reads mapping to the reference genome. Duplicate reads were
removed in order to avoid upward bias in the estimation of endogenous DNA content and to avoid inflation of length distribution for loci with deep coverage using a custom python script [20]. Endogenous DNA content ranged from 7.4% to 72.0% with a mean of 49.1% (Data S1A).

**Mitogenome data processing**

Reads were also mapped against the woolly rhinoceros mitochondrial reference genome (GenBank ID: FJ905813) using the above settings to generate mitochondrial BAM files for downstream processing. We imported the mitochondrial BAM files generated using SAMtools v1.3 as above to Geneious® v7.0.336 [57] where sequences were aligned using MUSCLE v3.8.31 [81]. We then called consensus sequences for positions with at least 5X coverage using a majority consensus rule, with ambiguous and low-coverage positions called as undetermined (N). Finally, we visually inspected the assembled sequences to assess overall coverage across the 16,438 base pairs (bp) of the mitogenome and quality of the SNPs identified.

**Nuclear genome data processing**

We selected one sample with 56.7% endogenous DNA and dated to 18,699-18,356 cal BP for deep-sequencing (ND035, Data S1A). We used the stock library described above as a template for PCR amplification and performed six indexing PCR reactions in order to increase library complexity. PCR amplification, purification and size selection were performed as described above and following Meyer & Kircher [45].

We used the same approach and parameters as described above (in “Endogenous DNA content estimation”) to generate BAM files, including mapping the woolly rhinoceros reads to the de-novo Sumatran Rhinoceros reference and duplicate read removal. Next, we used Picard v1.141 (http://broadinstitute.github.io/picard) to assign read group information including library, lane and sample identity to each bam file. Reads were then re-aligned around indels using GATK v3.4.0 [58]. Only reads/alignments with mapping quality ≥ 30 were kept for subsequent analysis. We estimated damage patterns and then performed a base-recalibration step on the BAM files using MapDamage v2.0 [59]. Finally, we estimated the depth of genome coverage using Qualimap v2.2.1 [60]. The average genome coverage was 13.6X with 70% of the genome with a genome coverage ≥ 10X (Data S1A). We then called variants using bcftools mpileup v1.3 [56] using a minimum depth of coverage of 1/3X of the average coverage and a maximum of 2X the average coverage, base quality ≥ 30 and removed SNPs within 5bp of indels. We also identified CpG sites using a custom script masking CG sites and removed them using BEDtools v2.29.2 [61]. Finally, for all downstream analyses, we excluded chromosome-linked scaffolds and masked repeat regions using BEDTools. Overall, we obtained 28,180,718 SNPs.

**Mitogenome data analysis**

Basic statistics including nucleotide diversity (π), number of haplotypes (n), haplotype diversity (d), and number of segregating sites (S) were performed using DnaSP6 v6.12.03 [62]. Nucleotide diversity (π) within the samples was 0.00268; the number of segregating sites (S) was 119; and haplotype diversity (d) was 0.989. Second, we created a median joining network in PopArt [63]. We added a traits block to the nexus alignment using a custom python script to visualize the samples based on geographic region (Figure S1A).

Third, we performed demographic reconstruction of woolly rhinoceros over the last 125 ky BP in BEAST v1.10.4 [8]. The evolutionary model for the 14 mitogenome dataset was determined to be HKY+I using the Bayesian Inference Criterion in jModelTest v2.1.9 [64]. Calibrated tip dates were added in BEAUti v1.10.4 [8] using the median value of dates listed in Data S1A. For one sample (ND045) that was dated at > 45 Cal ka BP (Data S1A), we used a prior with a wide boundary (uniform, initial value: 45300, lower: 0, upper: 500,000) since the rate of coalescence is inversely proportional to effective population size. We generated consensus sequences for all extant relative of the woolly rhinoceros [82]. All models were run using Beast v1.10.4 for 10 million generations with sampling every 1000 generations. We calculated marginal likelihoods for each tree model using path and stepping stone models implemented in Beast v1.10.4 (Data S1B). All output log files were visualized in Tracer v1.7.1 [66] in order to ensure convergence had occurred. Demographic reconstructions based on female effective population size for each tree model were also performed using Tracer (Figure S1B-d). Tree Annotator v1.10.4 [8] was used to remove 10% burn-in from the tree files. The phylogenies were then visualized in Figtree v1.4.4 [65].

**Demographic reconstruction**

We used the Pairwise Sequentially Markovian Coalescent (PSMC v0.6.5 [17]) model to infer the effective population sizes (N_e) of the woolly rhinoceros over time. This approach infers the distribution of the time to the most recent common ancestor (TMRCA) between the two alleles across all chromosomes using the density of heterozygous sites across the diploid genome of a single individual. Regions of low heterozygosity reflect recent coalescent events while regions of high heterozygosity reflect more ancient coalescent events. The rate of coalescent events in each segment is then informative about changes in effective population size through time since the rate of coalescence is inversely proportional to effective population size. We generated consensus sequences for all
autosomes of the ancient genome using the SAMtools v1.3 mpileup command and the ‘vcf2fq’ command from vcfutils.pl. We used filters for base quality, mapping quality and root-mean-squared mapping quality below 30, and depth below 1/3 and higher than 2-times the average coverage estimated for each specimen. In order to infer the distribution of the time to the most recent common ancestor (TMRCA) between the two copies of each chromosome from each individual across all autosomes, we set N (the number of iterations) = 25, t (Tmax) = 15 and p (atomic time interval) = 64 (4+25+2+4+6, for each of which parameters are estimated with 28 free interval parameters). We used a rate of $2.34 \times 10^{-8}$ substitutions/site/generation [12] and a generation time of 12 years [13]. Due to the uncertainty in substitution rates in Rhinocerotidae, we also tested rates of $1.2 \times 10^{-8}$ [83] and $3.3 \times 10^{-8}$ substitutions/site/generation based on horse [84].

**Heterozygosity and inbreeding**

We first estimated the overall autosomal heterozygosity using mlRho v2.7 [67] to estimate the population mutation rate (κ), which approximates expected heterozygosity under the infinite sites model. We filtered out bases with quality below 30, reads with mapping quality below 30 and positions with root-mean-squared mapping quality below 30. Because high or low coverage in some regions resulting from structural variation can create erroneous mapping to the reference genome and false heterozygous sites, we filtered out sites with depth lower than 1/3X and higher than 2X the estimated average coverage. The maximum likelihood approach implemented in mlRho has been shown to provide unbiased estimates of average within-individual heterozygosity at high coverage [67, 85].

Second, we estimated inbreeding by identifying runs of homozygosity (ROH) and calculating the individual inbreeding coefficients ($F_{ROH}$) with the sliding-window approach implemented in PLINK v1.9 [68]. We converted the filtered vcf file into a ped file and identified ROHs in all autosomal scaffolds. We used a sliding window size of 100 SNPs (–homozyg-window-snp 100). A window was then defined as homozygous if there were not more than 15 missing sites (homozyg-window-missing 15) and not more than 5 heterozygous sites per window (homozyg-window-het 5). If at least 5% of all windows that included a given SNP were defined as homozygous, the SNP was defined as being in a homozygous segment of a chromosome (homozyg-window-threshold 0.05). This threshold was chosen to ensure that the edges of a ROH are properly delimited. A homozygous segment was defined as a ROH if all of the following conditions were met: the segment included ≥25 SNPs (homozyg-snp 25) and covered ≥100 kb (homozyg-kb 100). Furthermore, the minimum SNP density was one SNP per 50 kb (homozyg-density 50) and the maximum distance between two neighboring SNPs was ≤1,000 kb (homozyg-gap 1,000). For the number of heterozygous sites within ROHs, we set the value at 750 (homozyg-het 750) in order to prevent sequencing errors to cut ROHs. We found that the majority of ROH were <0.5 Mb in length with a maximum ROH length of 2.5 Mb.

We also used a less stringent number of heterozygous sites per window of 10 (homozyg-window-het 10) to account for potential ancient DNA damage. However, results were very similar with the majority of ROH <0.5 Mb in length and maximum ROH of 3.4 Mb (Figure S2).

**Non-synonymous mutations**

We used SNPeff v4.3 [69] to annotate non-synonymous nucleotide substitutions in coding regions for the woolly rhinoceros. First, we generated a database for the Sumatran rhinoceros reference genome using the protein sequences extracted from our annotation. Second, we identified non-synonymous variants in two different SNPeff impact categories: (a) non-disruptive variants that might change protein function and effectiveness, hereafter referred to as missense variants; and (b) high, variants assumed to have high (disruptive) impact in the protein, probably causing protein truncation or triggering nonsense mediated decay (e.g., stop codons, splice donor variant and splice acceptor), hereafter referred to as loss of function (LoF) variants [69].

Next, we ranked all missense variants reported by SNPeff using relative change of selected physicochemical properties of substituted amino acids and using a custom script (https://github.com/NBISweden/simpred) (Data S1C). For every non-synonymous amino acid substitution, we computed three values reflecting the substitution-induced change in amino acid properties:

1. **aaIndex score.** Seven aaindex [36] descriptors of amino acid properties selected in [87, 88]. Every such descriptor represents a certain property of amino acid (e.g., its polarity as a numerical value corresponding to the magnitude of the property). These descriptors were chosen with biological interpretability in mind and span a 7-dimensional space (with every descriptor represented by one axis) that preserves full discernibility between any pair of amino acids. Now, we represented every amino acid as a point in this 7-dimensional space and following this step, we used multi-dimensional scaling (MDS) to find a low dimensional (2D) representation of the data points. In this low-dimensional representation, the simple Euclidean distance between any pair of points representing particular amino acids corresponds to the overall magnitude of change induced by substituting one amino acid from the pair to the other. This relation is naturally symmetric. We then constructed a relative substitution scores matrix where the score for a given pair of amino acids is relative to the maximal possible pairwise distance between a pair of amino acids.

2. **Sneath index** which takes into account various chemical properties of substituting amino acids [89]. Similarly to aaIndex, we use scores relative to maximal possible change.

3. **Non-exchangeability index**, which is an inverse of the exchangeability index proposed in Yampolsky and Stoltzfus [90]. The exchangeability index is based on a number of experimental studies and thus provides another perspective on potential amino
acid substitution impact. It is a non-symmetric index, i.e. non_exchgb(aa1 → aa2) ≠ non_exchgb(aa2 → aa1) and is not defined for all amino acid substitutions. We are also using relative non-exchangeability for consistency with the previously mentioned scores. For every type of score, value of 1.0 corresponds to the maximal possible change (impact) while low values indicate likely mild substitution effects.

Finally, we retained genes affected by missense and LoF variants and identified orthologs, and assessed the functional enrichment of these LoF variants using Panther with horse as the reference set [70] (Data S1D-H).