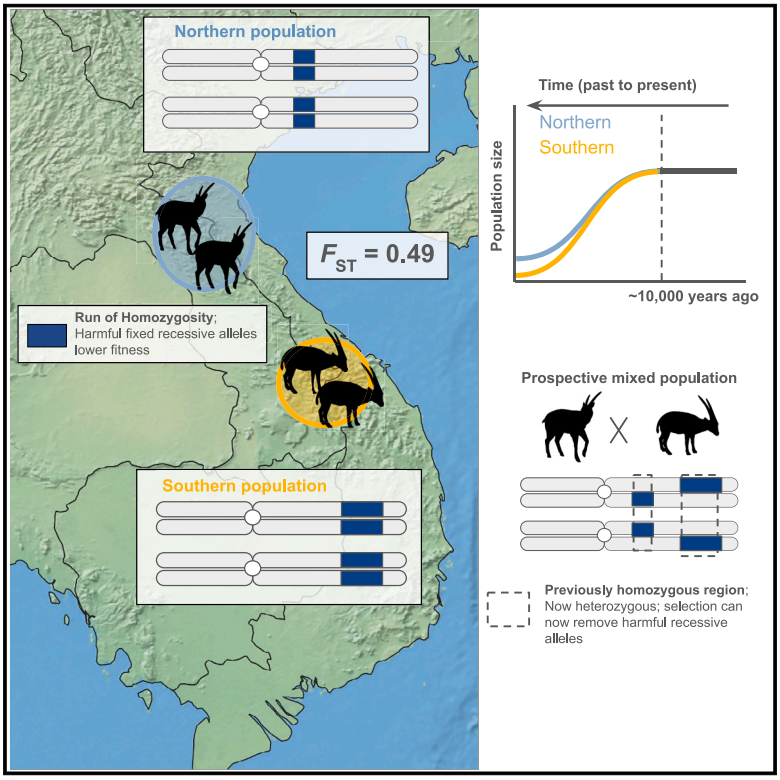
Cell

Genomes of critically endangered saola are shaped by population structure and purging

Graphical abstract



Highlights

- A reference genome and sequencing of 26 recently discovered and near-extinct saola
- Two highly differentiated populations diverged and gradually declined over >5,000 years
- Gradual decline caused extremely low genetic diversity and strong purging of genetic load
- Combining the two populations would reduce the otherwise high realized genetic load

Authors

Genís Garcia-Erill, Shanlin Liu, Minh Duc Le, ..., Anders Albrechtsen, M. Thomas P. Gilbert, Rasmus Heller

Correspondence

aalbrechtsen@bio.ku.dk (A.A.), tgilbert@sund.ku.dk (M.T.P.G.), rheller@bio.ku.dk (R.H.)

In brief

The generation of whole-genome sequencing data from 26 saolas provides insights into the phylogenetic placement, population structure, historical decline, and genetic diversity in one of the world's most mysterious and elusive large mammals.

Garcia-Erill et al., 2025, Cell *188*, 3102–3116
June 12, 2025
https://doi.org/10.1016/j.cell.2025.03.040

CellPress

CellPress
OPEN ACCESS

Cell

Article

Genomes of critically endangered saola are shaped by population structure and purging

Genís Garcia-Erill,[1,2,19] Shanlin Liu,[3,4,19] Minh Duc Le,[5,18] Martha M. Hurley,[6] Hung Dinh Nguyen,[7] Dzung Quoc Nguyen,[7] Dzung Huy Nguyen,[7] Cindy G. Santander,[1] Fátima Sánchez Barreiro,[4] Nuno Filipe Gomes Martins,[4] Kristian Hanghøj,[1] Faezah Mohd Salleh,[4,8] Jazmín Ramos-Madrigal,[4] Xi Wang,[1] Mikkel-Holger S. Sinding,[1] Hernán E. Morales,[4] Frederik Filip Stæger,[1] Nicholas Wilkinson,[9] Guanliang Meng,[10] Patrícia Pečnerová,[1] Chentao Yang,[11] Malthe Sebro Rasmussen,[1] Mikkel Schubert,[1] Robert R. Dunn,[12] Ida Moltke,[1] Guojie Zhang,[1,13] Lei Chen,[14] Wen Wang,[14] Trung Tien Cao,[15] Ha Manh Nguyen,[16] Hans R. Siegismund,[1] Anders Albrechtsen,[1,*] M. Thomas P. Gilbert,[4,17,*] and Rasmus Heller[1,20,*]

[1]Department of Biology, University of Copenhagen, Copenhagen, Denmark
[2]Bioinformatics Research Centre, Department of Molecular Biology and Genetics, Aarhus University, Aarhus, Denmark
[3]Key Laboratory of Zoological Systematics and Evolution, Institute of Zoology, Chinese Academy of Sciences, Beijing 100101, China
[4]Globe Institute, University of Copenhagen, Copenhagen, Denmark
[5]Faculty of Environmental Sciences, University of Science, Vietnam National University, Hanoi, 334 Nguyen Trai Road, Hanoi, Vietnam
[6]Center for Biodiversity and Conservation, American Museum of Natural History, New York, NY, USA
[7]Forest Inventory and Planning Institute, Ministry of Agriculture and Rural Development, Hanoi, Vietnam
[8]Department of Biosciences, Faculty of Science, Universiti Teknologi Malaysia, 81310 Johor Bahru, Johor, Malaysia
[9]Independent Consultant, Linton, Cambridgeshire, UK
[10]Zoological Research Museum Alexander Koenig, LIB, Bonn, Germany
[11]BGI Research, Shenzhen 518083, China
[12]Department of Applied Ecology, North Carolina State University, Raleigh, NC, USA
[13]Center of Evolutionary & Organismal Biology, Zhejiang University School of Medicine, Hangzhou 310058, China
[14]Center for Ecological and Environmental Science, Northwestern Polytechnical University, Xi'an 710072, China
[15]Institute of Biology, Chemistry and Environment, Vinh University, Vinh, Vietnam
[16]Center for Nature Conservation and Development, No. 05, 56/119 Tu Lien Street, Hanoi, Vietnam
[17]University Museum, NTNU, Trondheim, Norway
[18]Vietnam and Central Institute for Natural Resources and Environmental Studies, Vietnam National University, Hanoi, 19 Le Thanh Tong, Hanoi, Vietnam
[19]These authors contributed equally
[20]Lead contact
*Correspondence: aalbrechtsen@bio.ku.dk (A.A.), tgilbert@sund.ku.dk (M.T.P.G.), rheller@bio.ku.dk (R.H.)
https://doi.org/10.1016/j.cell.2025.03.040

SUMMARY

The saola is one of the most elusive large mammals, standing at the brink of extinction. We constructed a reference genome and resequenced 26 saola individuals, confirming the saola as a basal member of the Bovini. Despite its small geographic range, we found that the saola is partitioned into two populations with high genetic differentiation (F_{ST} = 0.49). We estimate that these populations diverged and started declining 5,000–20,000 years ago, possibly due to climate changes and exacerbated by increasing human activities. The saola has long tracts without genomic diversity; however, most of these tracts are not shared by the two populations. Saolas carry a high genetic load, yet their gradual decline resulted in the purging of the most deleterious genetic variation. Finally, we find that combining the two populations, e.g., in an eventual captive breeding program, would mitigate the genetic load and increase the odds of species survival.

INTRODUCTION

The saola (*Pseudoryx nghetinhensis*) is one of the rarest mammals in the world. This bovid was not scientifically described until 1993,[1] making it the most recently described large land mammal. The last large land mammal discovered before the saola was the kouprey (*Bos sauveli*) in 1937.[2,3] The saola is native to the forests of the Annamite mountain range along the border between Vietnam and Laos. It remains the only large terrestrial mammal that has yet to be observed alive by scientists in its natural habitat, despite the efforts of several research teams and conservation bodies (e.g., Long[4] and WWF-Vietnam[5]). Our very limited knowledge about this elusive animal is largely drawn from physical remains (primarily skulls and a few skins), anecdotes collected from inhabitants of the region, and five observations from camera-trapping surveys in Vietnam and Laos. According to the

International Union for Conservation of Nature (IUCN), it is Red Listed as critically endangered (CR). Its population size was assessed in 2015 to be 50–300 individuals, and its continued survival is in doubt.[6] The saola faces threats from indiscriminate snaring and disturbance and loss of its forest habitat.[7] Although over 20 individuals were captured alive by locals in the 1990s, attempts to keep them alive failed,[8] likely due to a lack of professional care. Today, a major effort is underway to build a well-equipped captive breeding facility for saola in Vietnam, in the hope that live individuals can be captured and transferred to the facility as a last chance to save it from extinction.[6] The plan is to use the captive population to reintroduce animals in a protected area where poaching is prevented.[9] Due to its low and continuously declining population size, the outlook for the saola is extremely precarious, and its evolutionary history and genetic diversity have yet to be studied because of the paucity of sample material.

Even if current efforts to locate and transfer saolas to protected facilities were successful, their long-term survival might still be challenged. Small populations are vulnerable to genomic erosion due to genetic drift and inbreeding, which lead to a loss of genetic diversity and an increase in harmful mutations, threatening their long-term survival.[10–13] However, in the last decade, theoretical and empirical studies have underscored that there is no simple relationship between the impact of demographic forces on genetic variation and extinction risk.[12,14,15] For example, long-term small population size tends to lead to the accumulation of weakly and moderately deleterious variation due to a reduced efficacy of selection,[16,17] but it also removes strongly deleterious variations through genetic purging.[18,19] Similarly, gene flow from external populations can improve population health by increasing genetic diversity and reducing the segregation of deleterious variation in what is known as genetic rescue.[20,21] However, gene flow can also increase extinction risk by introducing strongly deleterious mutations.[22,23] The effects of gene flow depend, among other factors, on whether the demographic history of the source population has allowed for purging of most of the strongly deleterious recessive mutations.[24] It is therefore paramount to have information about the demographic history, population structure, genetic diversity, and genetic load of an endangered species to inform potential management strategies.[25] However, due to the extremely limited genetic data from saola, these processes are entirely unknown in the species, which is a serious limitation for deciding appropriate conservation strategies.

In addition to the unresolved questions related to the species' genetic structure and conservation, the saola represents an evolutionary enigma, and its phylogenetic placement remains the outstanding phylogenetic conundrum for the Bovidae. Their ancestral bovid features, particularly the unusual combination of "caprine" and "bovine" morphological characteristics, led to contradicting propositions regarding its phylogenetic placement after its scientific description.[26,27] While cytogenetic and phylogenetic studies based on mtDNA and 13 introns have determined it is part of the tribe Bovini, its placement within the tribe differs in various analyses that have placed it inconsistently as a sister taxon to the Bovina,[28–30] the Bubalina,[31] or indeed as a basal branch within the Bovini tribe.[32] Whole-genome sequencing data can resolve such phylogenetic controversies[33] and, in doing so, also help to resolve the saola's evolutionary distinctiveness, a key component of its value to conservation.[34]

To obtain a better foundation for the management of the remaining genetic diversity in the species, we generated genomic data from a collection of 31 saola samples dating back to the early 1990s (Table S4). Sampling was restricted to scientific collections taken from animals that were previously deceased. We used these data to generate a draft reference genome for the saola and to perform phylogenomic and population genomic analyses of this critically endangered and extremely obscure mammal.

RESULTS

Generating a draft genome assembly to facilitate saola conservation genomics

Despite the relatively poor quality of the DNA that could be recovered from the best sample available (bioanalyzer visualized fragments <15 kb) (Figure S1A), we were able to generate a *de novo* genome sequence for a male saola. We built libraries of 250 bp, 500 bp, 2 kbp, and 5 kbp fragments (Table S1), sequenced with paired-end reads of 150 bp length to produce a draft genome assembly with scaffold N50 of 2.3 Mb using Platanus.[35] The assembly contained 5,952 out of 6,253 Benchmarking Universal Single-Copy Orthologs (BUSCO)[35,36] based on the Laurasiatheria (orthoDB v9) database (95.2%) (Figure S1B) and had a total length of 2.7 Gbp, consistent with the k-mer-based genome size estimate of 2.77 Gb.[37] With subsequent reference-assisted chromosome assembly (RACA)[38] *in silico* improvement, we were able to generate a total of 117 predicted chromosome fragments (PCFs), assembling 90% of the total genome (2.5 G) and improving the scaffold N50 to ca. 79 MB (Figure S1C). Nearly 49% of the genome is composed of repetitive elements, similar to that of cattle (48%),[39] the majority of which are long interspersed nuclear elements (LINEs; 26.22%) (Table S2). Using a combination of homology-based and *de novo* approaches, we annotated a total of 21,054 protein-coding genes (Table S3) and 7,512 non-coding functional RNAs. At 55.8%, the GC content of the coding region is considerably higher than that of the whole genome (41.9%), similar to reports of the GC distribution in other mammals.[40]

Phylogenomic analysis confirms the saola as a sister taxon to cattle and buffalo

A total of 51,644 non-overlapping sliding windows covering the genome with an average length of 50,875 bp were extracted and used for the phylogenetic tree inference, which placed the saola as a sister group to the cattle and water buffalo lineages (Figure 1). Using 4-fold degenerate sites, we subsequently estimated the divergence time of the saola lineage from the cattle/water buffalo lineage to ca. 14 million years ago (mya), in the middle of the Miocene epoch (Figure 1). The dominant tree topology was supported by an average of 67.9% of subtrees, with relatively low variability across the autosomes but with the X chromosome showing an elevated proportion (82.0%) of subtrees supporting the species tree. In the blocks with discordant

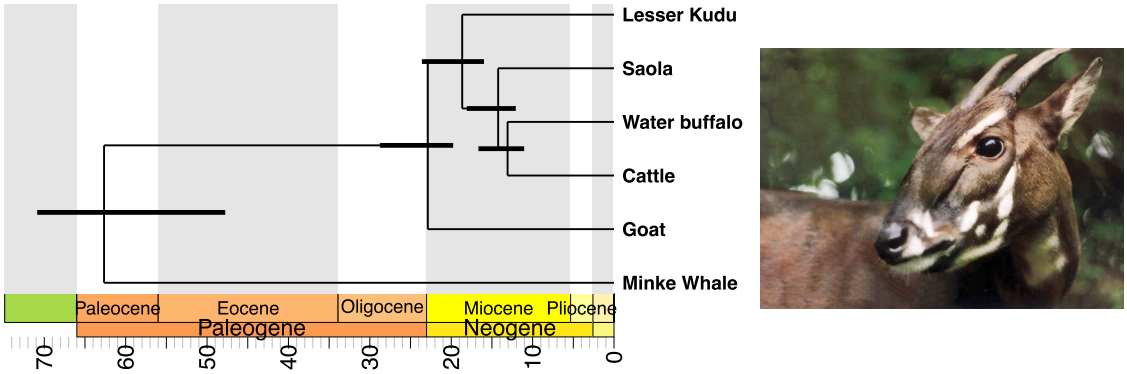


Figure 1. Phylogenetic analyses of the saola genome

Species tree of cattle, water buffalo, saola, goat, lesser kudu, and minke whale. The divergence time (in mya) was geologically scaled, and it is shown on each node as the posterior divergence time (thick black lines indicate 95% high posterior density). Next to the phylogeny is shown a photo of a saola, reproduced with permission from the copyright holder, ©Toon Fey/WWF.

See also Figure S1.

topologies, there was an overrepresentation of trees joining the saola with the water buffalo (27.9%), while a much smaller minority placed it with the cattle (4.5%) (Figure S1D).

Generating a population genomic dataset using conservative filtering of variants

We generated whole-genome resequencing data from 31 saola and, after removing two samples with insufficient endogenous DNA and merging three pairs of duplicated samples (Figure S2A; Table S4), we retained data from 26 individuals (Figures 2A and 2C). The samples showed high variability in sequencing depth (Figure 2C), DNA damage patterns (Figure S2B), and error rates (Figure S2C), necessitating the use of different analysis strategies and the inclusion of different sample subsets for each analysis (see STAR Methods; Figure S2D). Following strict filtering of sites to remove regions with low sequencing coverage, repetitive regions, and other regions likely to contain mapping and genotyping errors (Table S5), we retained a total of ≈1.2 Gb with a relatively uniform distribution across the genome (Figure S2E) for population genetic analyses, within which we identified 628,905 common (minor allele frequency [MAF] > 5%) transversion SNPs across the 26 saola samples. This underestimates the true number of common polymorphisms in the saola, as a large number of sites were conservatively removed to reduce errors, including all transition mutations, which typically account for approximately 2/3 of all SNPs.[41]

Two genetically differentiated saola populations with low genetic diversity and elevated proportion of deleterious genetic variation

A principal-component analysis, admixture analysis, and an mtDNA tree consistently split the samples into two clearly differentiated genetic groups concordant with their geographic origins and without evidence of further discrete substructure within them (Figures 2B, 2D, and S3A–S3C). We refer to them as the "northern" and "southern" populations in all subsequent analyses. The northern population is represented by 13 samples from Vũ Quang National Park in Vietnam, one sample from the Pù Mát National Park in Vietnam, and one sample from Nakai-Nam Theun National Park in Laos PDR, while the southern population is represented by one sample obtained in Huế City, seven from Đông Giang district, and four from Tây Giang District, Quang Nam Province, all in Vietnam (Figure 2A). Only one sample, which we labeled as N??1, clustered with the northern samples despite having a recorded southern origin (Table S4). On further investigation, it turned out that this sample was acquired from a souvenir shop in Huế City, and its true origin is unknown. As the illegal wildlife trade has been a concern in Vietnam over the last decades, it is possible that the saola horn on sale at the shop was trafficked from the northern range, a few hundred kilometers from Huế City. Therefore, and in light of the geographical distance and the genetic differentiation between the two regions, we consider it much more likely that the sample material was moved from its origin before being collected rather than it being a migrant, and we therefore grouped it with the northern population in all subsequent analyses. The clear separation into these two geographically consistent populations makes it highly unlikely that more samples are erroneously designated, despite the lack of comprehensive metadata for most samples.

The two-dimensional site frequency spectrum (2DSFS) between the two populations reveals that most of the segregating genetic variation in the saola is not shared between the two populations (Figure S3D), which results in an F_{ST} estimate of 0.47 and 0.49 depending on sample inclusion, mutation filtering criteria, and allele frequency inference methodologies (Table S6).

We calculated genome-wide heterozygosities for each individual to assess the levels of genetic diversity in the saola. In order to maximize the number of samples used while avoiding DNA damage biasing our results, we estimated heterozygosities using only transversion mutations and rescaled the estimates by a factor of 3. We used the three samples with an average sequencing depth >15× to support the heterozygosity estimation with different complementary methods (Figure S4A). We selected 8 samples with average depth above 6× that were not excessively affected by DNA damage and sequencing errors and considered them "high-quality" samples (Figures 2C, S2B, and S4B). For

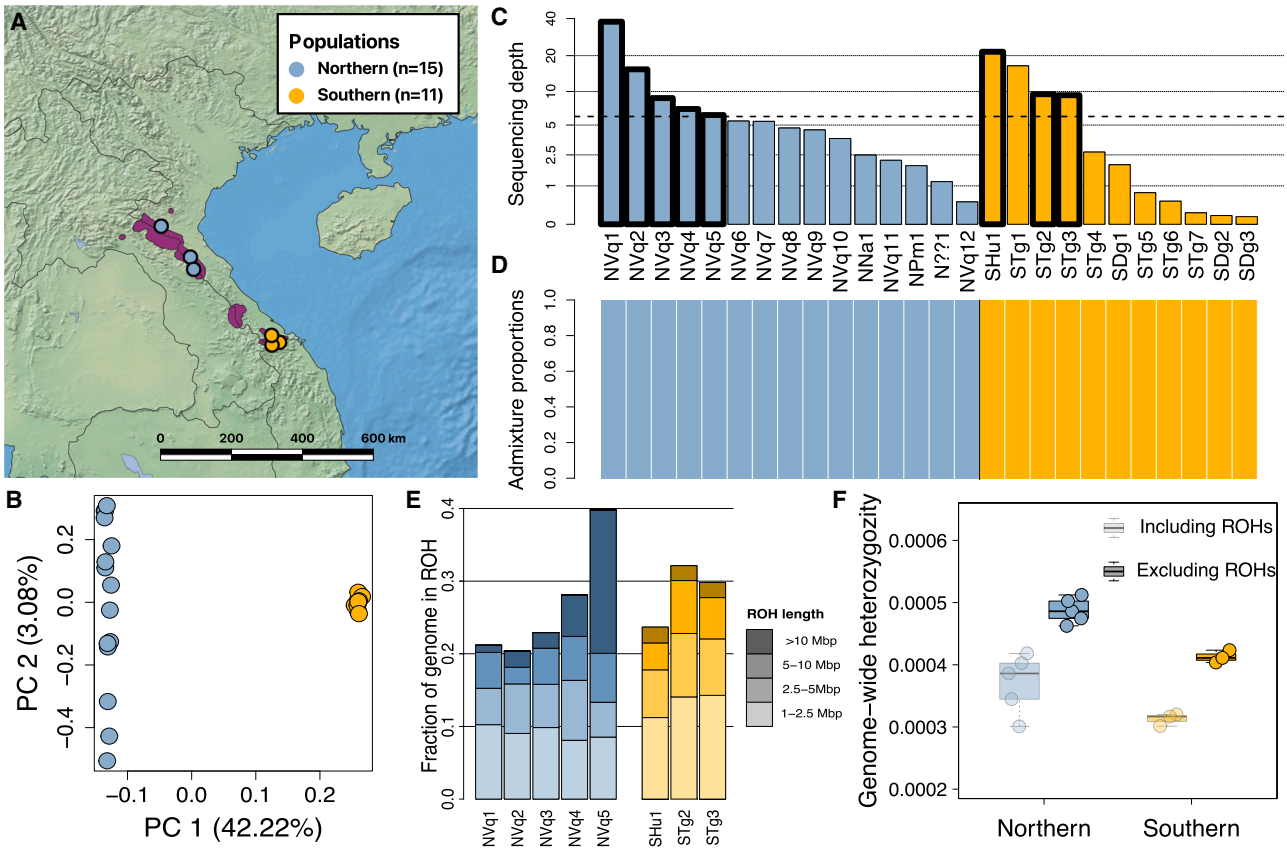


Figure 2. Population structure and genome-wide diversity

(A) Sampling map of the 26 saola individuals analyzed. The purple shading reflects the saola's distribution according to IUCN,[42] but the difficulty in detecting saola makes its actual distribution range highly uncertain, and therefore it does not necessarily represent the saola's current or historical range.

(B) The first two axes (45.36.2% of variation explained) of the principal-component analysis (PCA) are based on genotype likelihoods for the 26 individuals.

(C) Average sequencing depth of each sample, with the 8 high-quality samples highlighted. High-quality samples are those with an average sequencing depth larger than 6× (marked with a discontinuous line) and error rates in transversion mutations below 0.001 (see Figures S2C and S2D).

(D) Admixture proportions assuming K = 2.

(E) ROH for the high-quality samples highlighted in (C).

(F) Genome-wide heterozygosity estimates including and excluding regions in ROHs larger than 1 Mb for the same 8 high-quality samples shown in (C). Heterozygosity values are estimated using only transversion mutations and rescaled to account for all mutations (Figure S4A).

For sample IDs shown in (C)–(E), the first letter indicates inferred genetic population (N, northern; S, southern), while the second two letters indicate sampling locality (Vq: Vũ Quang National Park, Vietnam; Na: Nakai-Nam Theun National Park, Laos PDR; Pm: Pù Mát National Park, Vietnam; Hu: Huế, Vietnam; Tg: Tây Giang, Quang Nam, Vietnam; Dg: Đông Giang, Quang Nam, Vietnam;??: unknown).

See also Figures S2–S4.

these samples, it was possible to perform runs of homozygosity (ROH) estimation (Figure S4C), and we found that approximately 20%–40% of each saola genome was covered by long (>1 Mb) ROH, indicating recent inbreeding (Figure 2E). Samples from the southern population tended to have lower heterozygosity levels than those from the northern population, although the range of heterozygosity values for the two populations overlaps when ROHs are included (Figure 2F). Allele frequency-based estimates of inbreeding and kinship revealed some related pairs of individuals and four samples (three in the northern and one in the southern) with elevated inbreeding (Figure S4D), indicative of being the offspring of close relatives.

We assessed the saola's neutral and functional genetic variation by placing it in the context of previous estimates for other species. The saola showed remarkably low heterozygosity, at the lower end of previously published estimates from other species, and comparable with some other bovids believed to have been long restricted to small geographical areas (Figures 3A and S5A). Estimates of heterozygosity are sensitive to filtering decisions, and therefore cross-species comparisons are prone to noise, but qualitatively our conclusions regarding the low diversity of saola are robust to considerable noise in these estimates. To investigate the extent of potentially deleterious segregating genetic variation, we counted the non-synonymous variants and predicted loss-of-function (LOF) variants. To make the number comparable between species and to avoid issues with misspecification of the ancestral allele, we show the number of heterozygous putatively deleterious variants normalized by

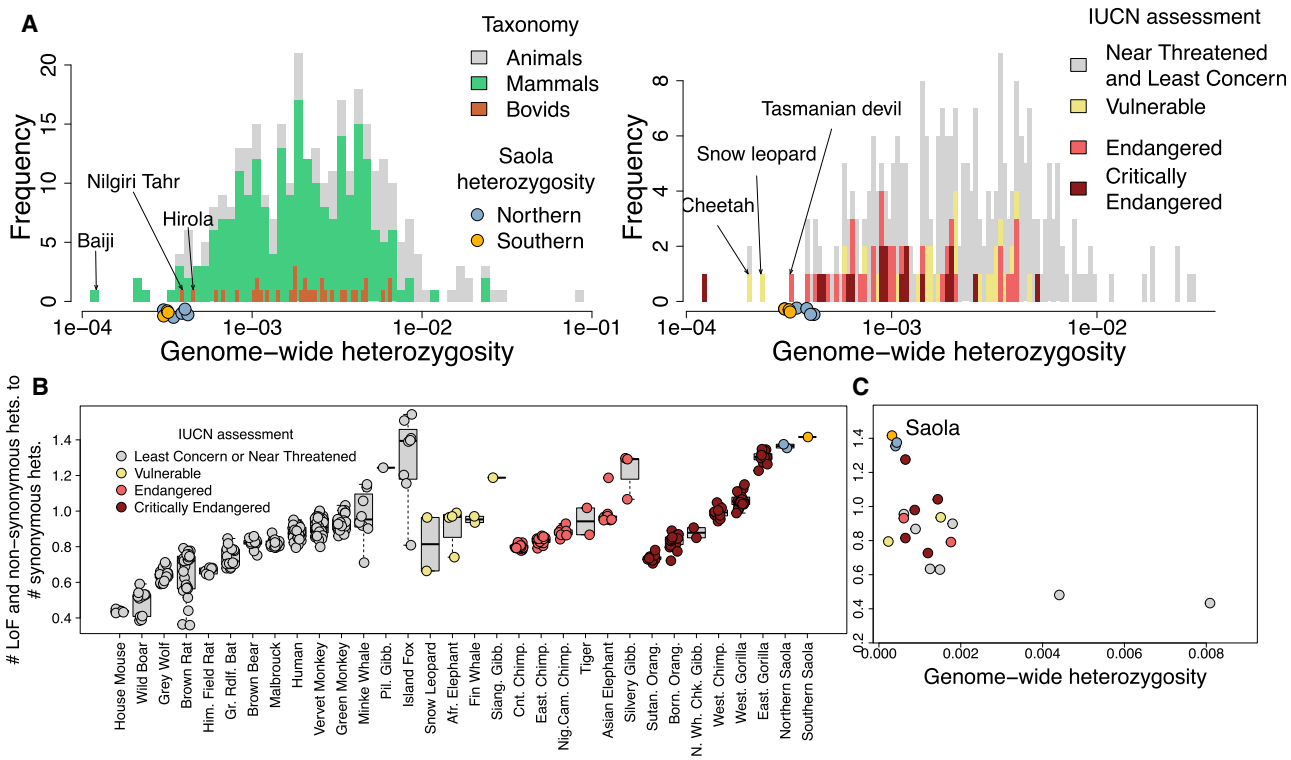


Figure 3. Saola genome-wide and functional genetic diversity in context

(A) Genome-wide heterozygosity values for the high-quality saola samples within the context of heterozygosity estimates for other animal species. In the left panel, the other species are highlighted by the relevant taxonomy at different levels. The right panel shows the subset of species with IUCN assessment colored by status.[45]

(B) Relative amount of segregating potentially deleterious genetic variation across different mammal species, measured as the ratio of the total number of coding heterozygous variants that impact the amino acid sequence (grouping together missense and LOF mutations) to the total number of heterozygous synonymous coding variants. The different species are grouped by their conservation status.

(C) Ratio of the number of potentially deleterious heterozygous sites (LOF and non-synomous) to synonymous heterozygous sites (same as in B) plotted against the estimate of genome-wide heterozygosity (same as shown in A), for those species for which there was data in both (A) and (B). See Figure S5C for the species each dot corresponds to. Species are colored by conservation status following the legend shown in (B).

See also Figure S5.

the heterozygous synonymous variants. Here, we observed a high relative amount of potentially deleterious variation segregating in saola, comparable only to estimates for eastern gorillas and island foxes (Figures 3B and S5B). When plotting the ratio of relative deleterious variation against the genome-wide heterozygosity, which is a proxy of effective population size, we found that saola and other species with lower genome-wide heterozygosity tend to have a higher ratio (Figures 3C and S5C). This pattern of segregating putatively deleterious variation is commonly interpreted to be caused by a reduction in the efficacy of selection and increased effect of genetic drift in small populations. However, we caution that there has been a recent dispute on whether changes in the efficacy of selection are needed to explain this pattern.[17,43,44]

Saola populations have retained complementary genetic diversity along the genome during their decline

We next sought to characterize the genomic landscape of diversity both within and between the two saola populations, with a special focus on regions devoid of diversity. We estimated the genetic diversity with Tajima's π in genomic windows within each of the two populations and within a combined population sample generated by pooling equal numbers of samples (three) from each population. In both of the populations, many windows across the genome showed no detectable genetic diversity, with the southern population exhibiting a higher number of depleted regions (Figure 4A). Genetic diversity in the combined population was higher, and the proportion of windows without detectable diversity was lower (Figures 4A and 4B). Next, we counted the number of 100 kb windows devoid of genetic diversity within and between populations. In most regions where both populations had no diversity, there was diversity in the combined population, indicating that each population was fixed for a different haplotype (Figure 4C). The pattern was consistent across window sizes from 50 to 500 kbp (Figure S6A) and was not driven by the bias introduced by missing sites in the estimation of genetic diversity (Figures S6B and S6C). Combining individuals from the two populations also increases the genetic diversity in regions where many samples are in an ROH (Figure 4D). Interestingly, we found that regions where ROHs accumulate

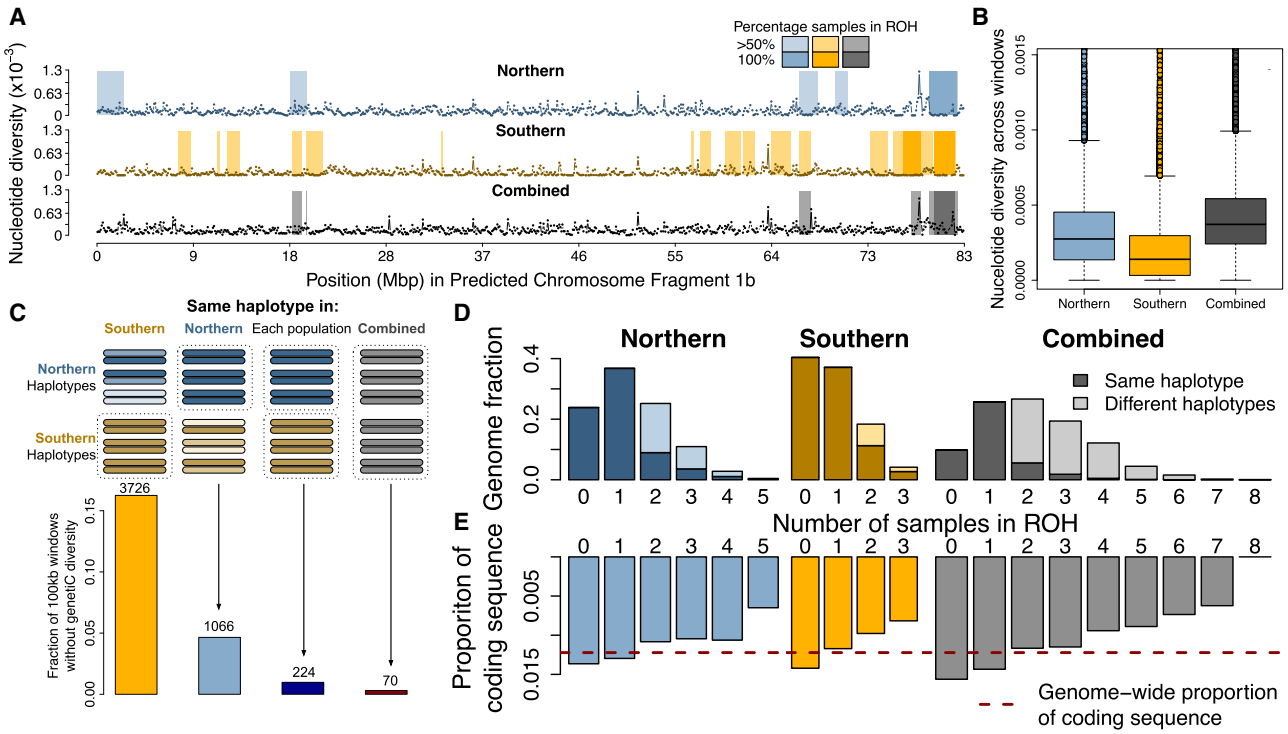


Figure 4. The genomic landscape of genetic diversity loss in saola

(A) Genetic diversity in 100 kb windows in an example Predicted Chromosome Fragment (PCF) for the northern population, the southern population, and a combined population where we mix samples from both populations. Genetic diversity π is estimated using three samples from each population and using only transversion mutations and rescaling by the expected 1:3 factor when using all mutations.

(B) Boxplot with the genome-wide distribution of nucleotide diversity in 100 kb windows for each of the three population groupings shown in (A).

(C) The intersection of windows without genetic diversity in the northern, southern, and combined populations for windows of 100 kb size. The y axis shows the proportion over all windows, while the absolute number of windows is shown above each bar.

(D) Fraction of the genome where samples in both populations or in each population are in ROH, colored depending on whether the ROH carries the same or different haplotypes.

(E) Proportion of coding sequence for each of the regions where a certain number of samples is in ROH.

See also Figure S6.

tend to have a lower proportion of coding sequences than the genome-wide average (Figure 4E). This pattern deviates significantly from the expectation if the accumulation of inbreeding tracts was random across the genome (Figure S6D). This pattern may result from the higher frequency of recessive deleterious variants in coding regions. Individuals with ROH in areas rich in coding regions are likely to experience reduced fitness, which promotes stronger genetic purging in these regions. Thereby, inbreeding tracts will preferentially accumulate in genomic regions with less coding sequence. Note that this does not mean that diversity is higher in coding regions but simply that ROHs are less prevalent there.

A long-term decline and recent split of the two saola populations

We used several complementary approaches to infer the saola's demographic history. We first inferred long-term effective population sizes (N_e) with the pairwise sequentially Markovian coalescent (PSMC) on the samples with depth >15×. Based on this, we found that N_e has been moderately low (starting at a maximum of ≈15,000) and continuously decreasing from around 0.5 mya, with a temporary recovery during the Late Pleistocene, followed by an accelerated decline 30 to 20 thousand years ago (kya). The estimated effective population size was never above 5,000 in the last 10,000 years (Figures 5A–5C and S7A). This last decline roughly coincides with the last glacial maximum (LGM) and is accompanied by a divergence of the N_e curve between the two populations (Figure 5A). Next, we used genomic regions where a sample from both the southern and northern populations was in ROH (around ≈120 Mbp for each of the two pairs of samples), and for which we therefore know the phased haplotypes. These regions were used to generate autosomal pseudodiploid samples and infer cross-population coalescence rates through time by running PSMC (Figure S7B). We then ran PSMC on a comparable region of the individual genomes (see STAR Methods; Figure S7B) to calculate the relative cross-coalescence rates through time, which revealed a decline in cross-coalescence between the populations coinciding with the visible separation of the population size trajectories during the LGM, reaching its minimum value around 2–5 kya (Figure 5A).

To explore the more recent dynamics of effective population size, we used population-level data from the northern population,

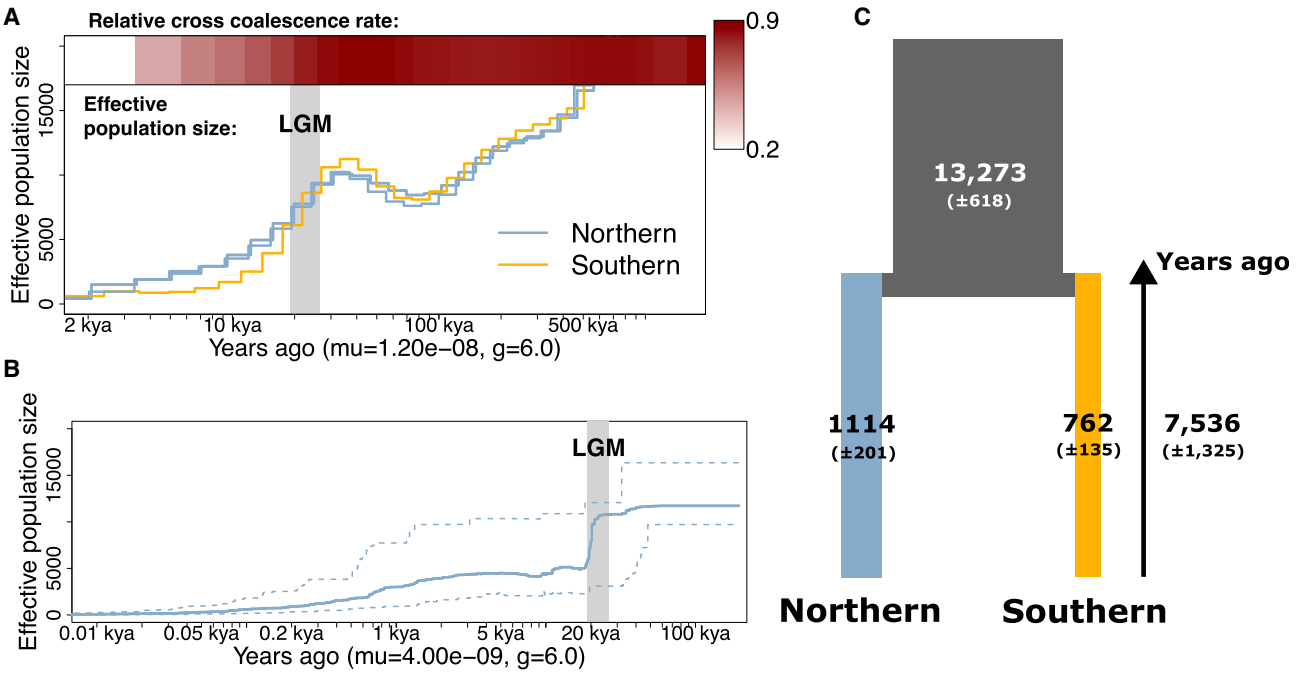


Figure 5. Demographic history

(A) Long-term effective population sizes for the two populations inferred with PSMC (bottom) and relative cross-coalescence rates through time estimated by using PSMC in regions where two samples from different populations are both in a ROH (top). All time intervals from 2 to 500 kya have more than 100 recombinations, which is above the minimum of 10 recommended for reliable PSMC inference.[46]

(B) Recent population size for the northern population inferred with stairway plot v2. The mutation rate is scaled to 1/3 of the genome-wide estimate because the SFS was estimated excluding all transition mutations. The gray shaded box in (A) and (B) indicates the time of the last glacial maximum (LGM, 19–26 kya).

(C) Joint demographic history of the two saola populations inferred with fastsimcoal v2 with the maximum likelihood parameter estimates and their corresponding jackknife standard errors in parentheses. Population sizes are not to scale.

See also Figure S7.

from which we had enough samples of relatively good quality to estimate the site frequency spectrum (SFS) (Figure S7B). This corroborated the population decline during the LGM and furthermore revealed that it was an abrupt decrease followed by a period of slow, but continuous, decline in population size (Figure 5B). The model also inferred a further recent decline in N_e to extremely low levels in the present ($N_e < 10$ diploid individuals). Finally, we investigated the population divergence process between the two populations by fitting a simple demographic model to the 2DSFS (Figures 5C and S3D). We recovered a very recent population divergence of 7.5 kya, which is roughly consistent with the time when PSMC relative cross-coalescence rates reach their minimum but not the time when they start decreasing (Figure 5A). Also consistent with other estimates, the demographic model suggested that divergence was accompanied by a 17- and 24-fold decline in the northern and southern population sizes, respectively. The high genetic differentiation between the two populations has therefore been driven by an extreme amount of drift during a short period of time due to low population sizes, rather than an old split time and more moderate drift.

Simulations show genetic purging and suggest population mixing will increase species viability

We sought to investigate how the inferred demography has impacted the accumulation of genetic load in the two saola populations by performing simulations of their complete coding sequence under the inferred demographic history. Our simulations were set up to answer two questions: (1) how would realized and total genetic load have accumulated in saolas over the inferred demographic history, allowing us to estimate how much purging took place, and (2) how will the accumulated genetic load impact population viability in prospective scenarios of a managed saola population? We used two different models for the distribution of dominance and selection coefficients (Figure S8A) that have previously been used for similar forward simulations of genetic load.[47,48] We decomposed the genetic load into contributions of different dominance coefficient bins. Simulations showed an overall increase in realized genetic load (the genetic load that is expressed as negative fitness effect; Bertorelle et al.[49]) during the saola population size decline and revealed that this is caused by an increased effect of recessive deleterious mutations due to the increased homozygosity when genetic drift increases (Figures 6A and S8B). However, using a measure of genetic purging based on the total genetic load in the population (including "masked" load not expressed[49]), we found that substantial genetic purging occurred under the inferred saola population history and is driven by selective elimination of the most deleterious recessive mutations, as evidenced in the much stronger decline of the load constituted by the most recessive and deleterious variants in the lower panel of Figure 6A. Hence,

although we show that realized load increased over the last 10,000 generations in the saola, purging removed much of the genetic load that the species would have otherwise accumulated. The two different models of selection and dominance coefficient distributions yielded different absolute values of genetic load, but the relative dynamics are highly similar, and the overall conclusion regarding the dynamics of genetic load and purging is robust to modeling assumptions (Figures S8B and S8C).

We furthermore tested several prospective scenarios to explore how demographic stochasticity and genetic load affect the relative viability of saola populations under a conservation scenario where a low number of individuals are maintained, emulating a captive breeding framework such as that proposed for the saola.[9] We performed non-Wright Fisher simulated scenarios where 4, 12, or 24 individuals from each population were used to found a captive breeding program, and comparable scenarios where a mixed captive breeding program was established with individuals from both populations. These simulations showed very low population viability in any scenario involving only 4 founding saolas of any origin (0–5 populations survived out of 200 simulations). Scenarios with 12 or 24 founding saolas had higher viability (21–60 and 70–145 populations survived out of 200 simulations), and we found that in these scenarios mixing of founders from both populations led to higher population viability (Figure 6B). Also, mixed-founder populations always had higher genetic diversity than non-mixed-founder populations (Figure 6B), emphasizing that mixed captive breeding programs would not only increase the odds of short-term species survival but also increase long-term evolutionary potential in the resulting population. In our simulations, realized genetic load in mixed-breeding programs tends to increase in the generations following the mixing, as deleterious variants from each population appear again in homozygous state, but the realized genetic load did not exceed that of the single population programs in the long term (Figure S9A). Mixed-founder breeding programs were superior regardless of the assumed model of selection and dominance coefficients, although the absolute viability of breeding programs differed substantially according to which model was assumed (Figures 6B and S9B). Therefore, predictions about viability based on population genetic simulations should be interpreted with caution, given their sensitivity to assumptions about unknown parameters.

To complement the above, we performed another series of captive breeding simulations but this time assumed a counterfactual retrospective demographic history in which the two saola populations declined abruptly rather than gradually (Figure S9C). As expected, purging was much less efficient under this demographic history. We then forward simulated captive breeding scenarios from this starting point and found that due to the less efficient purging, genetic load now led to much lower population viability across all scenarios (Figure S9C). This highlights that past demographic history has a crucial impact on the accumulation of genetic load and therefore on population viability.

DISCUSSION

We present the first genomic analyses of the saola, perhaps the world's most elusive large terrestrial mammal species. We confirm the placement of the saola as a basal Bovini, highlighting its evolutionary uniqueness as a lineage 14 million years divergent from other extant species. Hence, using the increased power of whole-genome data, we reject the phylogenetic positions inferred through previous analyses based upon cytogenetics[31] and single-marker mitochondrial genomes[29] and uphold the topology previously inferred from 13 intron regions,[32] and consistent with the assessment of some previous studies.[50]

We show that saolas are partitioned into two highly genetically differentiated populations. The relatively recent inferred onset of divergence and population decline at ≈20 kya coincides with likely changes in the distribution of dense forest habitat during and after the LGM,[51–54] suggesting that these major events in saola population history could be driven by habitat fragmentation and isolation into two geographically distinct areas. Similar fragmentation has also been invoked to explain the phylogeography of co-distributed species such as the large-antlered muntjac.[55] Our demographic modeling estimates that this previously unknown saola population divergence may not have been complete until around ≈5 kya, indicating that the isolation of the two populations may have been a gradual process that occurred over millennia. The transition of human societies from a hunter-gatherer to an agricultural lifestyle, which occurred around 4 kya in present-day Vietnam,[56,57] could have contributed to the genetic isolation between two saola populations through an expansion of human activities, such as hunting, burning, and rice cultivation, and associated forest losses in the area. The recent decline in saola population sizes hence appears to be the continuation of a millennia-old decline, possibly instigated by a combination of climate change and early human activities.

We show that a gradual population decline has allowed the saola populations to eliminate some of the most highly deleterious mutations by purging and thus reduce the risk of inbreeding depression in the current populations relative to what would have occurred if the decline had happened more abruptly. This is in line with previous research on small and endangered populations, e.g., mountain gorillas,[19,58] island foxes,[59] vaquita,[60,61] kākāpō,[62] and the Ethiopian wolf.[63] Our simulations suggest that purging has removed almost two-thirds of the total genetic load present at the onset of the saola demographic decline. However, it is important to note that our analyses also show that the saola nonetheless carries an elevated relative amount of segregating potentially deleterious mutations when compared with most other species, which is expected due to a reduction in the efficiency of selection in small populations.[16] Rather than being mutually contradictory, these two observations are consistent with recent findings regarding the typical outcome of reduced population sizes. In small populations, genetic drift leads to an increase in the mildly and slightly deleterious mutations, while purging reduces overall genetic load by removing the strongly deleterious recessive mutations that contribute the most to inbreeding depression.[14] We show that the relative amount of deleterious variation segregating in saola is comparable to that of island foxes and eastern gorillas, two species that have also purged deleterious variations due to historically low population sizes.[19,59]

The discovery of two highly genetically differentiated populations has a potentially profound impact on saola management

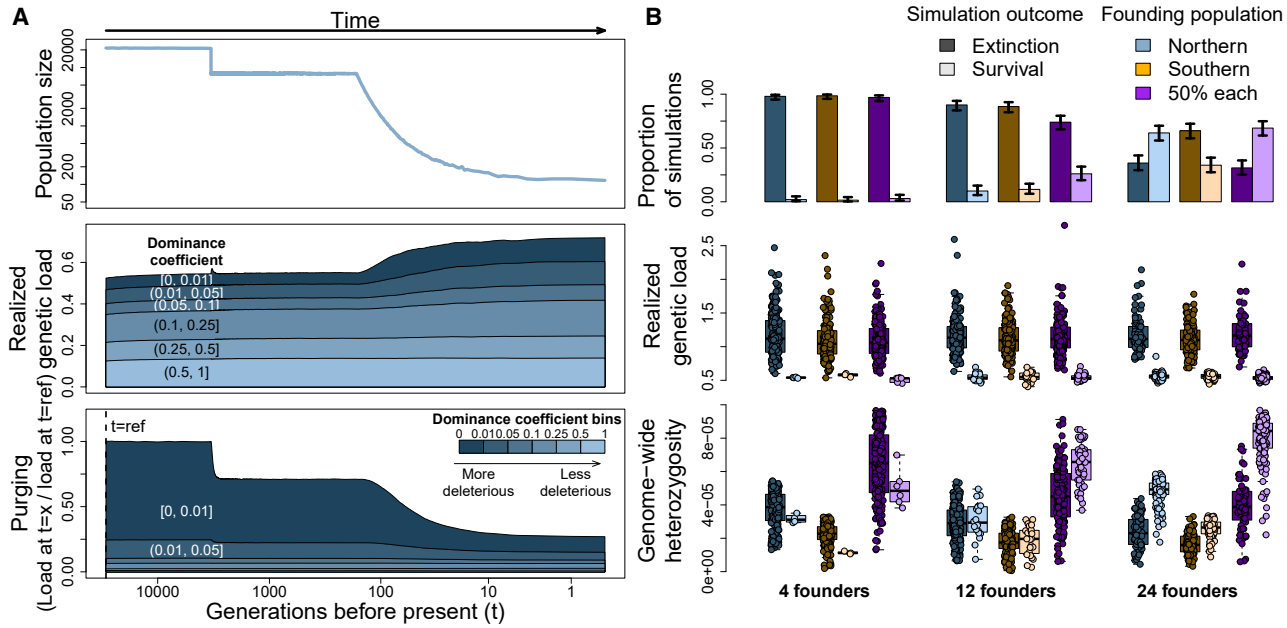


Figure 6. Simulation of genetic load in the saola

(A) Trajectories of realized genetic load (middle panel) and purging of genetic load (bottom panel), for the northern population simulated under the inferred saola demography (upper panel). Each dominance coefficient (h) category contains the cumulative value of the genetic load for all variants with a certain dominance value, such that the first category ($h \geq 0$ in darker colors) contains the total value of the corresponding genetic load measure across all mutations. Purging was quantified as the genetic load divided by the genetic load level at the time of the population split (see STAR Methods for details and Figure S8B for the trajectories of genetic load without normalization and for the corresponding genetic load trajectories for the southern population).

(B) Viability of saola populations under future conservation scenarios varying the number of founders and the population source of the founders. The upper panel shows the fraction of 200 simulations under each scenario where a captive population survived or went extinct, with error bars indicating the binomial standard error. The middle and bottom panel show the realized genetic load and genome-wide heterozygosity, respectively, at the last generation for each simulation separated by whether the population survived or went extinct. Simulations are done using the Pérez-Pereira model of fitness and dominance coefficients; see Figures S8C and S9B for the corresponding results under the Kyriazis model.

See also Figures S8 and S9.

strategies. We show that the isolation into two populations has led to the retention of somewhat complementary genetic variation in the species as a whole. The populations have diverged very recently from an evolutionary point of view, and the high differentiation has been driven by substantial drift due to low population sizes. Therefore, and due to the similarity of the wet evergreen forest habitat in the ranges of the two populations, it is unlikely that the two populations have developed any significant local adaptations or genetic incompatibilities as a result of their divergence. As a result, conserving the two populations separately is unlikely to conserve unique adaptive variation between them. Moreover, we show by simulations that offspring resulting from interbreeding between the two populations would have higher fitness due to the reduction in genetic load, a phenomenon referred to as genetic rescue.[64,65] While genetic rescue as a management strategy has been controversial due to both the risk of outbreeding depression[66–68] and the resulting genetic homogenization, which may be perceived as "unnatural,"[69,70] there is a growing realization that assisted gene flow is often beneficial for a species.[65,71–75] We also demonstrate that the positive outcome of genetic rescue in saola is contingent on the inferred demographic history characterized by a gradual population decline and that a more sudden decline would have led to other conclusions. Hence, we highlight that conservation strategies depend on information about past demographic history, and we demonstrate how simulations of load are a vital tool for assessing genetic rescue and other conservation regimes.[62,76] Current hopes for the survival of the saola are pinned on the establishment of a captive population,[9] but there are no living individuals of known location.[77] Placing our results in this light, the unambiguous conservation recommendation is to manage the saola as a single population, maximizing genetic diversity by integrating the genetic heritage of both the northern and southern populations, regardless of whether a captive breeding population can be established, and despite the high genetic differentiation between them. This would increase the saola's long-term survival chances by both reducing genetic load and increasing the amount of standing genetic variation available to adapt to future environmental changes. If a captive breeding program becomes feasible, genomic data should be actively incorporated to optimize the retention of genetic diversity and minimize load.[78] We emphasize, however, that our study only addresses the optimal management of the genetic resources, and it is outside the scope of our study to evaluate whether captive breeding is in fact the optimal conservation strategy for the saola. We also caution against using our analyses to conclude whether a

captive breeding strategy would succeed or fail and emphasize that genetic load may not be the most urgent threat facing the saola.

The present status of the saola is precarious. There are legitimate concerns that it may already be extinct in the wild.[79] However, assuming there are still some individuals surviving in the wild, the genomic dataset generated here should help improve new non-invasive DNA-based monitoring techniques such as screening for saola DNA from leech blood meals[80] or using environmental DNA.[81] These will be vital in locating surviving individuals for an eventual captive breeding program, as well as for inferring its geographic range, home ranges, and other unknown life history traits. In less optimistic scenarios where the saola goes extinct, our data can provide valuable information by providing a snapshot of the genetic variation segregating in the species at a time—roughly 1990s—when it was present in higher numbers than today.[77] This information can potentially be used in de-extinction efforts, where it is important to have knowledge about naturally occurring genetic variation.[82] De-extinction, once widely rejected for the saola,[8] is now being openly discussed among organizations involved in the most recent attempts to locate surviving saola in Vietnam (B. Long, personal communication). Ultimately, we hope that the data and inferences presented here will contribute to accelerating and supporting ongoing efforts for *in situ* protection and assist the design of an *ex situ* captive management program, thus facilitating the re-establishment of a saola population in the wild.

Limitations of the study

The fact that the available saola samples contained sub-optimal quality DNA means there is a risk that damage-driven miscoding lesions may affect our results. To mitigate this, we took extreme care in considering the potential for errors or artifacts in the sequencing data when conducting analyses and only used a subset of samples that proved reliable and were sequenced to higher depth for the analyses that are most sensitive, e.g., ROH estimation and PSMC. Wherever possible, we corroborated results by using alternative and/or complementary methods, e.g., in the heterozygosity and F_{ST} estimation. We additionally acknowledge that our forward-looking simulations do not incorporate all relevant elements of biological complexity, such as social structure or mating behavior, as we do not have access to any relevant data to parameterize these, nor is any such information likely to be forthcoming. Our results should therefore solely be interpreted as a quantification of how genetic load and demographic stochasticity (i.e., the possibility of fixation of one of the sexes) would progress in the population under the assumptions of the specific past demographic history and already accumulated genetic load. Finally, we acknowledge and emphasize that any management of the genetic resources in saola is contingent on the ability to locate live animals in the wild or on the establishment of future de-extinction programs.

RESOURCE AVAILABILITY

Lead contact

Requests for further information and resources should be directed and will be fulfilled by the lead contact, Rasmus Heller (rheller@bio.ku.dk).

Materials availability

This study did not generate new, unique reagents.

Data and code availability

- The saola reference genome has been deposited to NCBI, and the sequencing data generated in this study is available in Sequence Read Archive (SRA), both under BioProject ID PRJNA688353.
- Code used for the analyses is available at https://github.com/GenisGE/saola.

ACKNOWLEDGMENTS

We thank Amal al-Chaer for assistance with the handling of samples in the lab. We are very grateful to Andrew Tilker for his input on the saola conservation situation and various other valuable insights. We are very indebted to the people involved in the scientific discovery and the collection of sample material of the saola and to those involved in subsequent saola conservation efforts, including but not limited to Vu Van Dung, Pham Mong Giao, Nguyen Ngoc Chinh, Do Tuoc, Peter Actander, John MacKinnon, Eleanor Sterling, Bill Robichaud, and WWF Asia. We also thank Sarah Mak, Shyam Gopalakrishnan, Filipe G. Vieira, Marcela Sandoval Velasco, José Alfredo Samaniego, Tomàs Marquès-Bonet, Marc de Manuel, Juraj Bergman, and Mengjun Wu for their technical assistance and comments throughout this study. Binia De Cahsan Westbury is thanked for her help designing artwork for the study. Three anonymous reviewers are thanked for their helpful comments, which improved the manuscript. The research reported in this manuscript was funded by the Vietnamese Ministry of Science and Technology's Program 562 (grant no. ĐTĐL. CN-64/19) to M.D.L. S.L. was supported by scientific projects from the Institute of Zoology, the Chinese Academy of Sciences (grant no. 2023IOZ0104), the National Natural Science Foundation of China (32470455), and the Ministry of Science and Technology of the People's Republic of China (2023YFD2201804). The research reported in this manuscript was also funded by the Vietnamese Ministry of Science and Technology's Program 562 (grant no. ĐTĐL. CN-64/19) to M.D.L. Sample collection was supported in part by a grant to M.M.H. from the Disney Wildlife Conservation Fund. M.-H.S.S. was supported by The Carlsberg Foundation (CF20-0355). M.S. was supported by the Novo Nordisk Foundation (NNF23SA0084103). A.A. was funded by the Novo Nordisk Foundation (NNF20OC0061343). M.T.P.G. was supported by the Danish National Research Foundation (DNRF143) and a European Research Council Consolidator Grant (ERC CoG 681396 "Extinction Genomics"). R.H. was supported by a Villum Young Investigator grant (VKR023447), an Independent Research Fund Denmark Sapere Aude grant (DFF8049-00098B), and a European Research Council Starting Grant (ERC-2018-STG-804679).

AUTHOR CONTRIBUTIONS

Conceptualization, M.D.L., N.D.H., N.Q.D., N.H.D., G.Z., W.W., T.C.T., H.R.S., A.A., M.T.P.G., and R.H.; methodology, G.G.-E., S.L., C.G.S., F.S.B., N.F.G.M., K.H., F.M.S., J.R.-M., X.W., M.-H.S.S., F.F.S., P.P., M.S.R., M.S., L.C., I.M., A.A., M.T.P.G., and R.H.; software, G.G.-E., S.L., C.G.S., K.H., M.S.R., and M.S.; validation, G.G.-E., S.L., C.G.S., F.S.B., N.F.G.M., F.M.S., J.R.-M., M.-H.S.S., H.E.M., M.S., F.F.S., P.P., A.A., M.T.P.G., and R.H.; formal analysis, G.G.-E., S.L., C.G.S., K.H., X.W., F.F.S., P.P., M.S.R., and L.C.; investigation, S.L., F.S.B., N.F.G.M., K.H., F.M.S., J.R.-M., M.-H.S.S., and M.S.; resources, M.D.L., M.M.H., N.D.H., N.Q.D., N.H.D., W.W., G.Z., A.A., M.T.P.G., and R.H.; data curation, G.G.-E., S.L., N.D.H., N.Q.D., N.H.D., M.D.L., K.H., N.W., and M.S.; writing – original draft, G.G.-E., S.L., M.T.P.G., and R.H.; writing – review and editing, all authors; visualization, G.G.-E., S.L., and F.F.S.; supervision, I.M., W.W., A.A., M.T.P.G., and R.H.; project administration, A.A., M.T.P.G., and R.H.; and funding acquisition, S.L., M.D.L., I.M., W.W., R.H., A.A., and M.T.P.G.

DECLARATION OF INTERESTS

The authors declare no competing interests.

STAR★METHODS

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

- KEY RESOURCES TABLE
- EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS
- METHOD DETAILS
  - DNA extraction and sequencing library preparation
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Reference genome assembly and annotation
  - Phylogenetic placement of the saola
  - Mapping of resequenced samples
  - Reference genome masking for population genomic analyses
  - Ancestral state estimation in the saola reference genome
  - DNA damage patterns and error rate estimation
  - Mitochondrial genome reconstruction
  - Mitochondrial tree
  - Population SNP calling and genotype likelihood estimation
  - Genotype calling
  - Duplicates, relatedness and inbreeding inference
  - Inference of population structure
  - Estimation of heterozygosity
  - Inference of runs of homozygosity
  - Population site frequency spectrum estimation
  - F_{ST} estimation
  - Comparison of saola genetic diversity with other species
  - Genetic diversity in windows across the genome
  - Sharing of regions in ROHs across samples
  - Demographic history analyses
  - Simulations of genetic load

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.cell.2025.03.040.

Received: September 25, 2024
Revised: December 20, 2024
Accepted: March 25, 2025
Published: April 21, 2025

REFERENCES

1. Van Dung, V., Giao, P.M., Chinh, N.N., Tuoc, D., Arctander, P., and MacKinnon, J. (1993). A new species of living bovid from Vietnam. Nature *363*, 443–445. https://doi.org/10.1038/363443a0.
2. MacKinnon, J.R., and Stuart, S.N. (1989). The Kouprey: An Action Plan for Its Conservation (IUCN).
3. Urbain, A. (1939). Note complémentaire sur le Bœuf sauvage du Cambodge (*Bos (Bibos) sauveli Urbain*). Bull. Mus. Natl. Hist. Nat. *2*, 519–520.
4. Long, B. (2018). Can the Elusive Saola Be Saved from Extinction? (Capeia).
5. WWF-Vietnam (2024). Wildlife of the Central Annamites (World Wildlife).
6. IUCN (2021). Position Statement on the Conservation of Saola (IUCN).
7. Kemp, N., Dilger, M., Burgess, N., and Van Dung, C. (1997). The saola *Pseudoryx nghetinhensis* in Vietnam – new information on distribution and habitat preferences, and conservation needs. Oryx *31*, 37–45.
8. Stone, R. (2006). Wildlife conservation. The saola's last stand. Science *314*, 1380–1383. https://doi.org/10.1126/science.314.5804.1380.
9. Tilker, A., Long, B., Gray, T.N.E., Robichaud, W., Van Ngoc, T., Vu Linh, N., Holland, J., Shurter, S., Comizzoli, P., Thomas, P., et al. (2017). Saving the saola from extinction. Science *357*, 1248. https://doi.org/10.1126/science.aap9591.
10. Caughley, G. (1994). Directions in Conservation Biology. J. Animal Ecol. *63*, 215–244. https://doi.org/10.2307/5542.
11. Bijlsma, R., and Loeschcke, V. (2012). Genetic erosion impedes adaptive responses to stressful environments. Evol. Appl. *5*, 117–129. https://doi.org/10.1111/j.1752-4571.2011.00214.x.
12. Díez-Del-Molino, D., Sánchez-Barreiro, F., Barnes, I., Gilbert, M.T.P., and Dalén, L. (2018). Quantifying Temporal Genomic Erosion in Endangered Species. Trends Ecol. Evol. *33*, 176–185. https://doi.org/10.1016/j.tree.2017.12.002.
13. van Oosterhout, C., Speak, S.A., Birley, T., Bortoluzzi, C., Percival-Alwyn, L., Urban, L.H., Groombridge, J.J., Segelbacher, G., and Morales, H.E. (2022). Genomic erosion in the assessment of species extinction risk and recovery potential. Preprint at bioRxiv. https://doi.org/10.1101/2022.09.13.507768.
14. Dussex, N., Morales, H.E., Grossen, C., Dalén, L., and van Oosterhout, C. (2023). Purging and accumulation of genetic load in conservation. Trends Ecol. Evol. *38*, 961–969. https://doi.org/10.1016/j.tree.2023.05.008.
15. Robinson, J.A., Kyriazis, C.C., Yuan, S.C., and Lohmueller, K.E. (2023). Deleterious Variation in Natural Populations and Implications for Conservation Genetics. Annu. Rev. Anim. Biosci. *11*, 93–114. https://doi.org/10.1146/annurev-animal-080522-093311.
16. Ohta, T. (1992). The Nearly Neutral Theory of Molecular Evolution. Annu. Rev. Ecol. Syst. *23*, 263–286. https://doi.org/10.1146/annurev.es.23.110192.001403.
17. Leroy, T., Rousselle, M., Tilak, M.-K., Caizergues, A.E., Scornavacca, C., Recuerda, M., Fuchs, J., Illera, J.C., De Swardt, D.H., Blanco, G., et al. (2021). Island songbirds as windows into evolution in small populations. Curr. Biol. *31*, 1303–1310.e4. https://doi.org/10.1016/j.cub.2020.12.040.
18. Kirkpatrick, M., and Jarne, P. (2000). The Effects of a Bottleneck on Inbreeding Depression and the Genetic Load. Am. Nat. *155*, 154–167. https://doi.org/10.1086/303312.
19. Xue, Y., Prado-Martinez, J., Sudmant, P.H., Narasimhan, V., Ayub, Q., Szpak, M., Frandsen, P., Chen, Y., Yngvadottir, B., Cooper, D.N., et al. (2015). Mountain gorilla genomes reveal the impact of long-term population decline and inbreeding. Science *348*, 242–245. https://doi.org/10.1126/science.aaa3952.
20. Miller, J.M., Poissant, J., Hogg, J.T., and Coltman, D.W. (2012). Genomic consequences of genetic rescue in an insular population of bighorn sheep (*Ovis canadensis*). Mol. Ecol. *21*, 1583–1596. https://doi.org/10.1111/j.1365-294X.2011.05427.x.
21. Whiteley, A.R., Fitzpatrick, S.W., Funk, W.C., and Tallmon, D.A. (2015). Genetic rescue to the rescue. Trends Ecol. Evol. *30*, 42–49. https://doi.org/10.1016/j.tree.2014.10.009.
22. Hedrick, P.W., Robinson, J.A., Peterson, R.O., and Vucetich, J.A. (2019). Genetics and extinction and the example of Isle Royale wolves. Animal Conservation *22*, 302–309. https://doi.org/10.1111/acv.12479.
23. Robinson, J.A., Räikkönen, J., Vucetich, L.M., Vucetich, J.A., Peterson, R.O., Lohmueller, K.E., and Wayne, R.K. (2019). Genomic signatures of extensive inbreeding in Isle Royale wolves, a population on the threshold of extinction. Sci. Adv. *5*, eaau0757. https://doi.org/10.1126/sciadv.aau0757.
24. Kyriazis, C.C., Wayne, R.K., and Lohmueller, K.E. (2021). Strongly deleterious mutations are a primary determinant of extinction risk due to inbreeding depression. Evol. Lett. *5*, 33–47. https://doi.org/10.1002/evl3.209.
25. Hohenlohe, P.A., Funk, W.C., and Rajora, O.P. (2021). Population genomics for wildlife conservation and management. Mol. Ecol. *30*, 62–82. https://doi.org/10.1111/mec.15720.
26. Robichaud, W.G. (1998). Physical and Behavioral Description of a Captive Saola, *Pseudoryx nghetinhensis*. Journal of Mammalogy *79*, 394–405. https://doi.org/10.2307/1382970.
27. Thomas, H. (1994). Anatomie crânienne et relations phylogénétiques du nouveau bovidé (*Pseudoryx nghetinhensis*) découvert dans la cordillère annamitique au Vietnam. Mammalia *58*, 453–482. https://doi.org/10.1515/mamm.1994.58.3.453.

28. Gatesy, J., and Arctander, P. (2000). Hidden morphological support for the phylogenetic placement of *Pseudoryx nghetinhensis* with bovine bovids: a combined analysis of gross anatomical evidence and DNA sequences from five genes. Syst. Biol. *49*, 515–538. https://doi.org/10.1080/10635159950127376.

29. Hassanin, A., Delsuc, F., Ropiquet, A., Hammer, C., Jansen van Vuuren, B., Matthee, C., Ruiz-Garcia, M., Catzeflis, F., Areskoug, V., Nguyen, T.T., et al. (2012). Pattern and timing of diversification of Cetartiodactyla (Mammalia, Laurasiatheria), as revealed by a comprehensive analysis of mitochondrial genomes. C. R. Biol. *335*, 32–50. https://doi.org/10.1016/j.crvi.2011.11.002.

30. Hassanin, A., and Douzery, E.J. (1999). Evolutionary affinities of the enigmatic saola (*Pseudoryx nghetinhensis*) in the context of the molecular phylogeny of Bovidae. Proc. Biol. Sci. *266*, 893–900. https://doi.org/10.1098/rspb.1999.0720.

31. Nguyen, T.T., Aniskin, V.M., Gerbault-Seureau, M., Planton, H., Renard, J.P., Nguyen, B.X., Hassanin, A., and Volobouev, V.T. (2008). Phylogenetic position of the saola (*Pseudoryx nghetinhensis*) inferred from cytogenetic analysis of eleven species of Bovidae. Cytogenet. Genome Res. *122*, 41–54. https://doi.org/10.1159/000151315.

32. Hassanin, A., An, J., Ropiquet, A., Nguyen, T.T., and Couloux, A. (2013). Combining multiple autosomal introns for studying shallow phylogeny and taxonomy of Laurasiatherian mammals: Application to the tribe Bovini (Cetartiodactyla, Bovidae). Mol. Phylogenet. Evol. *66*, 766–775. https://doi.org/10.1016/j.ympev.2012.11.003.

33. Jarvis, E.D., Mirarab, S., Aberer, A.J., Li, B., Houde, P., Li, C., Ho, S.Y.W., Faircloth, B.C., Nabholz, B., Howard, J.T., et al. (2014). Whole-genome analyses resolve early branches in the tree of life of modern birds. Science *346*, 1320–1331. https://doi.org/10.1126/science.1253451.

34. Isaac, N.J.B., Turvey, S.T., Collen, B., Waterman, C., and Baillie, J.E.M. (2007). Mammals on the EDGE: conservation priorities based on threat and phylogeny. PLoS One *2*, e296. https://doi.org/10.1371/journal.pone.0000296.

35. Kajitani, R., Toshimoto, K., Noguchi, H., Toyoda, A., Ogura, Y., Okuno, M., Yabana, M., Harada, M., Nagayasu, E., Maruyama, H., et al. (2014). Efficient de novo assembly of highly heterozygous genomes from whole-genome shotgun short reads. Genome Res. *24*, 1384–1395. https://doi.org/10.1101/gr.170720.113.

36. Simão, F.A., Waterhouse, R.M., Ioannidis, P., Kriventseva, E.V., and Zdobnov, E.M. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics *31*, 3210–3212. https://doi.org/10.1093/bioinformatics/btv351.

37. Marçais, G., and Kingsford, C. (2011). A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. Bioinformatics *27*, 764–770. https://doi.org/10.1093/bioinformatics/btr011.

38. Kim, J., Larkin, D.M., Cai, Q., Asan, Z., Zhang, Y., Ge, R.-L., Auvil, L., Capitanu, B., Zhang, G., Lewin, H.A., et al. (2013). Reference-assisted chromosome assembly. Proc. Natl. Acad. Sci. USA *110*, 1785–1790. https://doi.org/10.1073/pnas.1220349110.

39. Elsik, C.G., Tellam, R.L., Worley, K.C., Gibbs, R.A., Muzny, D.M., Weinstock, G.M., Adelson, D.L., Eichler, E.E., Elnitski, L., et al.; Bovine; Genome; Sequencing; Analysis Consortium (2009). The genome sequence of taurine cattle: a window to ruminant biology and evolution. Science *324*, 522–528. https://doi.org/10.1126/science.1169588.

40. Jiang, Y., Xie, M., Chen, W., Talbot, R., Maddox, J.F., Faraut, T., Wu, C., Muzny, D.M., Li, Y., Zhang, W., et al. (2014). The sheep genome illuminates biology of the rumen and lipid metabolism. Science *344*, 1168–1173. https://doi.org/10.1126/science.1252806.

41. 1000 Genomes Project Consortium, Abecasis, G.R., Altshuler, D., Auton, A., Brooks, L.D., Durbin, R.M., Gibbs, R.A., Hurles, M.E., and McVean, G.A. (2010). A map of human genome variation from population-scale sequencing. Nature *467*, 1061–1073. https://doi.org/10.1038/nature09534.

42. IUCN (International Union for Conservation of Nature) (2008). *Pseudoryx nghetinhensis*. The IUCN Red List of Threatened Species Version 2024-1.

43. Kratochvíl, L., and Rovatsos, M. (2022). Ratios can be misleading for detecting selection. Curr. Biol. *32*, R28–R30. https://doi.org/10.1016/j.cub.2021.11.066.

44. Leroy, T., and Nabholz, B. (2022). Response to Kratochvíl and Rovatsos. Curr. Biol. *32*, R30–R31. https://doi.org/10.1016/j.cub.2021.11.067.

45. IUCN (2021). The IUCN Red List of Threatened Species. Version 2021. https://www.iucnredlist.org.

46. Li, H., and Durbin, R. (2011). Inference of human population history from individual whole-genome sequences. Nature *475*, 493–496. https://doi.org/10.1038/nature10231.

47. Kyriazis, C.C., Robinson, J.A., and Lohmueller, K.E. (2022). Using computational simulations to quantify genetic load and predict extinction risk. Preprint at bioRxiv. https://doi.org/10.1101/2022.08.12.503792.

48. Pérez-Pereira, N., Caballero, A., and García-Dorado, A. (2021). Reviewing the consequences of genetic purging on the success of rescue programs. Conserv. Genet. *23*, 1–17.

49. Bertorelle, G., Raffini, F., Bosse, M., Bortoluzzi, C., Iannucci, A., Trucchi, E., Morales, H.E., and van Oosterhout, C. (2022). Genetic load: genomic estimates and applications in non-model animals. Nat. Rev. Genet. *23*, 492–503. https://doi.org/10.1038/s41576-022-00448-x.

50. Robichaud, W., and Timmins, R. (2004). The natural history of saola (*Pseudoryx nghetinhensis*) and the species' distribution in Laos. In Rediscovering the Saola, J. Hardcastle, S. Cox, N.T. Dao, and A.G. Johns, eds. (WWF Indochina Programme), pp. 27–28.

51. Wurster, C.M., Bird, M.I., Bull, I.D., Creed, F., Bryant, C., Dungait, J.A.J., and Paz, V. (2010). Forest contraction in north equatorial Southeast Asia during the Last Glacial Period. Proc. Natl. Acad. Sci. USA *107*, 15508–15511. https://doi.org/10.1073/pnas.1005507107.

52. Suraprasit, K., Shoocongdej, R., Chintakanon, K., and Bocherens, H. (2021). Late Pleistocene human paleoecology in the highland savanna ecosystem of mainland Southeast Asia. Sci. Rep. *11*, 16756. https://doi.org/10.1038/s41598-021-96260-4.

53. Morley, R.J., and Morley, H.P. (2022). 1 - The prelude to the Holocene: tropical Asia during the Pleistocene. In Holocene Climate Change and Environment, N. Kumaran and P. Damodara, eds. (Elsevier), pp. 1–32.

54. Woodruff, D.S. (2010). Biogeography and conservation in Southeast Asia: how 2.7 million years of repeated environmental fluctuations affect today's patterns and the future of the remaining refugial-phase biodiversity. Biodivers. Conserv. *19*, 919–941. https://doi.org/10.1007/s10531-010-9783-3.

55. Stimpson, C.M., Utting, B., O'Donnell, S., Huong, N.T.M., Kahlert, T., Manh, B.V., Khanh, P.S., and Rabett, R.J. (2019). An 11 000-year-old giant muntjac subfossil from Northern Vietnam: implications for past and present populations. R. Soc. Open Sci. *6*, 181461. https://doi.org/10.1098/rsos.181461.

56. Fuller, D.Q., and Cobo Castillo, C.C. (2021). 4 The origins and spread of cereal agriculture in Mainland Southeast Asia. In A Comprehensive Guide, P. Sidwell and M. Jenny, eds. (De Gruyter), pp. 45–60. https://doi.org/10.1515/9783110558142-004.

57. Jones, R.K., Piper, P.J., Wood, R., Nguyen, A.T., and Oxenham, M.F. (2019). The Neolithic transition in Vietnam: Assessing evidence for early pig management and domesticated dog. J. Archaeol. Sci.: Rep. *28*, 102042. https://doi.org/10.1016/j.jasrep.2019.102042.

58. van der Valk, T., Díez-Del-Molino, D., Marques-Bonet, T., Guschanski, K., and Dalén, L. (2019). Historical Genomes Reveal the Genomic Consequences of Recent Population Decline in Eastern Gorillas. Curr. Biol. *29*, 165–170.e6. https://doi.org/10.1016/j.cub.2018.11.055.

59. Robinson, J.A., Brown, C., Kim, B.Y., Lohmueller, K.E., and Wayne, R.K. (2018). Purging of Strongly Deleterious Mutations Explains Long-Term Persistence and Absence of Inbreeding Depression in Island Foxes. Curr. Biol. *28*, 3487–3494.e4. https://doi.org/10.1016/j.cub.2018.08.066.

60. Robinson, J.A., Kyriazis, C.C., Nigenda-Morales, S.F., Beichman, A.C., Rojas-Bracho, L., Robertson, K.M., Fontaine, M.C., Wayne, R.K., Lohmueller, K.E., Taylor, B.L., et al. (2022). The critically endangered vaquita is not doomed to extinction by inbreeding depression. Science *376*, 635–639. https://doi.org/10.1126/science.abm1742.

61. Morin, P.A., Archer, F.I., Avila, C.D., Balacco, J.R., Bukhman, Y.V., Chow, W., Fedrigo, O., Formenti, G., Fronczek, J.A., Fungtammasan, A., et al. (2021). Reference genome and demographic history of the most endangered marine mammal, the vaquita. Mol. Ecol. Resour. *21*, 1008–1020. https://doi.org/10.1111/1755-0998.13284.

62. Dussex, N., van der Valk, T., Morales, H.E., Wheat, C.W., Díez-Del-Molino, D., von Seth, J., Foster, Y., Kutschera, V.E., Guschanski, K., Rhie, A., et al. (2021). Population genomics of the critically endangered kākāpō. Cell Genom. *1*, 100002. https://doi.org/10.1016/j.xgen.2021.100002.

63. Mooney, J.A., Marsden, C.D., Yohannes, A., Wayne, R.K., and Lohmueller, K.E. (2023). Long-term Small Population Size, Deleterious Variation, and Altitude Adaptation in the Ethiopian Wolf, a Severely Endangered Canid. Mol. Biol. Evol. *40*, msac277. https://doi.org/10.1093/molbev/msac277.

64. Tallmon, D.A., Luikart, G., and Waples, R.S. (2004). The alluring simplicity and complex reality of genetic rescue. Trends Ecol. Evol. *19*, 489–496. https://doi.org/10.1016/j.tree.2004.07.003.

65. Bell, D.A., Robinson, Z.L., Funk, W.C., Fitzpatrick, S.W., Allendorf, F.W., Tallmon, D.A., and Whiteley, A.R. (2019). The Exciting Potential and Remaining Uncertainties of Genetic Rescue. Trends Ecol. Evol. *34*, 1070–1079. https://doi.org/10.1016/j.tree.2019.06.006.

66. Templeton, A.R. (1986). Coadaptation and outbreeding depression. In Conservation Biology: the Science of Scarcity and Diversity, M.E. Soulé, ed. (Sinauer Associates), pp. 105–116.

67. Edmands, S. (2007). Between a rock and a hard place: evaluating the relative risks of inbreeding and outbreeding for conservation and management. Mol. Ecol. *16*, 463–475. https://doi.org/10.1111/j.1365-294X.2006.03148.x.

68. Frankham, R., Ballou, J.D., Ralls, K., Eldridge, M., Dudash, M.R., Fenster, C.B., Lacy, R.C., and Sunnucks, P. (2017). Genetic Management of Fragmented Animal and Plant Populations (Oxford University Press) https://doi.org/10.1093/oso/9780198783398.001.0001.

69. Love Stowell, S.M., Pinzone, C.A., and Martin, A.P. (2017). Overcoming barriers to active interventions for genetic diversity. Biodivers. Conserv. *26*, 1753–1765. https://doi.org/10.1007/s10531-017-1330-z.

70. Kolodny, O., McLaren, M.R., Greenbaum, G., Ramakrishnan, U., Feldman, M.W., Petrov, D., and Taylor, R.W. (2019). Reconsidering the management paradigm of fragmented populations. Preprint at bioRxiv. https://doi.org/10.1101/649129.

71. Johnson, W.E., Onorato, D.P., Roelke, M.E., Land, E.D., Cunningham, M., Belden, R.C., McBride, R., Jansen, D., Lotz, M., Shindle, D., et al. (2010). Genetic restoration of the Florida panther. Science *329*, 1641–1645. https://doi.org/10.1126/science.1192891.

72. Frankham, R., Ballou, J.D., Eldridge, M.D.B., Lacy, R.C., Ralls, K., Dudash, M.R., and Fenster, C.B. (2011). Predicting the probability of outbreeding depression. Conserv. Biol. *25*, 465–475. https://doi.org/10.1111/j.1523-1739.2011.01662.x.

73. Chan, W.Y., Hoffmann, A.A., and van Oppen, M.J.H. (2019). Hybridization as a conservation management tool. Conserv. Lett. *12*, e12652. https://doi.org/10.1111/conl.12652.

74. Hasselgren, M., Angerbjörn, A., Eide, N.E., Erlandsson, R., Flagstad, Ø., Landa, A., Wallén, J., and Norén, K. (2018). Genetic rescue in an inbred Arctic fox (*Vulpes lagopus*) population. Proc. Biol. Sci. *285*, 20172814. https://doi.org/10.1098/rspb.2017.2814.

75. Pavlova, A., Schneller, N.M., Lintermans, M., Beitzel, M., Robledo-Ruiz, D.A., and Sunnucks, P. (2024). Planning and implementing genetic rescue of an endangered freshwater fish population in a regulated river, where low flow reduces breeding opportunities and may trigger inbreeding depression. Evol. Appl. *17*, e13679. https://doi.org/10.1111/eva.13679.

76. Femerling, G., van Oosterhout, C., Feng, S., Bristol, R.M., Zhang, G., Groombridge, J., P Gilbert, M.T., and Morales, H.E. (2023). Genetic Load and Adaptive Potential of a Recovered Avian Species that Narrowly Avoided Extinction. Mol. Biol. Evol. *40*, msad256. https://doi.org/10.1093/molbev/msad256.

77. Timmins, R.J., Hedges, S., and Robichaud, W. (2020). *Pseudoryx nghetinhensis* (amended version of 2016 assessment). The IUCN Red List of Threatened Species eT18597A46364962. https://dx.doi.org/10.2305/IUCN.UK.2020-1.RLTS.T18597A166485696.en.

78. Speak, S.A., Birley, T., Bortoluzzi, C., Clark, M.D., Percival-Alwyn, L., Morales, H.E., and van Oosterhout, C. (2024). Genomics-informed captive breeding can reduce inbreeding depression and the genetic load in zoo populations. Mol. Ecol. Resour. *24*, e13967. https://doi.org/10.1111/1755-0998.13967.

79. Saola Foundation (2022). Annual Report 2022 (Saola Foundation for Annamite Mountains Conservation).

80. Schnell, I.B., Thomsen, P.F., Wilkinson, N., Rasmussen, M., Jensen, L.R.D., Willerslev, E., Bertelsen, M.F., and Gilbert, M.T.P. (2012). Screening mammal biodiversity using DNA from leeches. Curr. Biol. *22*, R262–R263. https://doi.org/10.1016/j.cub.2012.02.058.

81. Thomsen, P.F., and Willerslev, E. (2015). Environmental DNA – An emerging tool in conservation for monitoring past and present biodiversity. Biol. Conserv. *183*, 4–18.

82. Robert, A., Thévenin, C., Princé, K., Sarrazin, F., and Clavel, J. (2017). De-extinction and evolution. Funct. Ecol. *31*, 1021–1031.

83. Das, A., Panitz, F., Gregersen, V.R., Bendixen, C., and Holm, L.-E. (2015). Deep sequencing of Danish Holstein dairy cattle for variant detection and insight into potential loss-of-function variants in protein coding genes. BMC Genomics *16*, 1043. https://doi.org/10.1186/s12864-015-2249-y.

84. Liu, B., Yuan, J., Yiu, S.-M., Li, Z., Xie, Y., Chen, Y., Shi, Y., Zhang, H., Li, Y., Lam, T.-W., et al. (2012). COPE: an accurate k-mer-based pair-end reads connection tool to facilitate genome assembly. Bioinformatics *28*, 2870–2874. https://doi.org/10.1093/bioinformatics/bts563.

85. Kelley, D.R., Schatz, M.C., and Salzberg, S.L. (2010). Quake: quality-aware detection and correction of sequencing errors. Genome Biol. *11*, R116. https://doi.org/10.1186/gb-2010-11-11-r116.

86. Xu, H., Luo, X., Qian, J., Pang, X., Song, J., Qian, G., Chen, J., and Chen, S. (2012). FastUniq: A Fast De Novo Duplicates Removal Tool for Paired Short Reads. PLoS One *7*, e52249. https://doi.org/10.1371/journal.pone.0052249.

87. Kuhn, R.M., Haussler, D., and Kent, W.J. (2013). The UCSC genome browser and associated tools. Brief. Bioinform. *14*, 144–161. https://doi.org/10.1093/bib/bbs038.

88. Smit Arian, F.A., Hubley, R., and Green, P. (2013). RepeatMasker Home Page. http://www.repeatmasker.org/.

89. Stanke, M., and Morgenstern, B. (2005). AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. Nucleic Acids Res. *33*, W465–W467. https://doi.org/10.1093/nar/gki458.

90. Majoros, W.H., Pertea, M., and Salzberg, S.L. (2004). TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders. Bioinformatics *20*, 2878–2879. https://doi.org/10.1093/bioinformatics/bth315.

91. Haas, B.J., Salzberg, S.L., Zhu, W., Pertea, M., Allen, J.E., Orvis, J., White, O., Buell, C.R., and Wortman, J.R. (2008). Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. Genome Biol. *9*, R7. https://doi.org/10.1186/gb-2008-9-1-r7.

92. Kiełbasa, S.M., Wan, R., Sato, K., Horton, P., and Frith, M.C. (2011). Adaptive seeds tame genomic sequence comparison. Genome Res. *21*, 487–493. https://doi.org/10.1101/gr.113985.110.

93. Blanchette, M., Kent, W.J., Riemer, C., Elnitski, L., Smit, A.F.A., Roskin, K.M., Baertsch, R., Rosenbloom, K., Clawson, H., Green, E.D., et al. (2004). Aligning multiple genomic sequences with the threaded blockset aligner. Genome Res. *14*, 708–715. https://doi.org/10.1101/gr.1933104.

94. Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics *30*, 1312–1313. https://doi.org/10.1093/bioinformatics/btu033.

95. Mirarab, S., Reaz, R., Bayzid, M.S., Zimmermann, T., Swenson, M.S., and Warnow, T. (2014). ASTRAL: genome-scale coalescent-based species tree estimation. Bioinformatics *30*, i541–i548. https://doi.org/10.1093/bioinformatics/btu462.

96. Schubert, M., Ermini, L., Der Sarkissian, C., Jónsson, H., Ginolhac, A., Schaefer, R., Martin, M.D., Fernández, R., Kircher, M., McCue, M., et al. (2014). Characterization of ancient and modern genomes by SNP detection and phylogenomic and metagenomic analysis using PALEOMIX. Nat. Protoc. *9*, 1056–1082.

97. Schubert, M., Lindgreen, S., and Orlando, L. (2016). AdapterRemoval v2: rapid adapter trimming, identification, and read merging. BMC Res. Notes *9*, 88. https://doi.org/10.1186/s13104-016-1900-2.

98. Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics *25*, 1754–1760. https://doi.org/10.1093/bioinformatics/btp324.

99. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R.; 1000 Genome Project Data Processing Subgroup (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078–2079. https://doi.org/10.1093/bioinformatics/btp352.

100. Pockrandt, C., Alzamel, M., Iliopoulos, C.S., and Reinert, K. (2020). GenMap: ultra-fast computation of genome mappability. Bioinformatics *36*, 3687–3692. https://doi.org/10.1093/bioinformatics/btaa222.

101. Nursyifa, C., Brüniche-Olsen, A., Garcia-Erill, G., Heller, R., and Albrechtsen, A. (2022). Joint identification of sex and sex-linked scaffolds in non-model organisms using low depth sequencing data. Mol. Ecol. Resour. *22*, 458–467. https://doi.org/10.1111/1755-0998.13491.

102. Korneliussen, T.S., Albrechtsen, A., and Nielsen, R. (2014). ANGSD: Analysis of Next Generation Sequencing Data. BMC Bioinformatics *15*, 356. https://doi.org/10.1186/s12859-014-0356-4.

103. Meisner, J., and Albrechtsen, A. (2018). Inferring Population Structure and Admixture Proportions in Low-Depth NGS Data. Genetics *210*, 719–731. https://doi.org/10.1534/genetics.118.301336.

104. Jónsson, H., Ginolhac, A., Schubert, M., Johnson, P.L.F., and Orlando, L. (2013). mapDamage2.0: fast approximate Bayesian estimates of ancient DNA damage parameters. Bioinformatics *29*, 1682–1684.

105. Katoh, K., and Standley, D.M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol. Biol. Evol. *30*, 772–780. https://doi.org/10.1093/molbev/mst010.

106. Darriba, D., Posada, D., Kozlov, A.M., Stamatakis, A., Morel, B., and Flouri, T. (2020). ModelTest-NG: A New and Scalable Tool for the Selection of DNA and Protein Evolutionary Models. Mol. Biol. Evol. *37*, 291–294. https://doi.org/10.1093/molbev/msz189.

107. Bouckaert, R., Vaughan, T.G., Barido-Sottani, J., Duchêne, S., Fourment, M., Gavryushkina, A., Heled, J., Jones, G., Kühnert, D., De Maio, N., et al. (2019). BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis. PLoS Comput. Biol. *15*, e1006650. https://doi.org/10.1371/journal.pcbi.1006650.

108. Rambaut, A., Drummond, A.J., Xie, D., Baele, G., and Suchard, M.A. (2018). Posterior Summarization in Bayesian Phylogenetics Using Tracer 1.7. Syst. Biol. *67*, 901–904. https://doi.org/10.1093/sysbio/syy032.

109. Danecek, P., Bonfield, J.K., Liddle, J., Marshall, J., Ohan, V., Pollard, M.O., Whitwham, A., Keane, T., McCarthy, S.A., Davies, R.M., et al. (2021). Twelve years of SAMtools and BCFtools. GigaScience *10*, giab008. https://doi.org/10.1093/gigascience/giab008.

110. Rasmussen, M.S., Garcia-Erill, G., Korneliussen, T.S., Wiuf, C., and Albrechtsen, A. (2022). Estimation of site frequency spectra from low-coverage sequencing data using stochastic EM reduces overfitting, runtime, and memory usage. Genetics *222*, iyac148. https://doi.org/10.1093/genetics/iyac148.

111. Hanghøj, K., Moltke, I., Andersen, P.A., Manica, A., and Korneliussen, T.S. (2019). Fast and accurate relatedness estimation from high-throughput sequencing data in the presence of inbreeding. GigaScience *8*, giz034. https://doi.org/10.1093/gigascience/giz034.

112. Skotte, L., Korneliussen, T.S., and Albrechtsen, A. (2013). Estimating individual admixture proportions from next generation sequencing data. Genetics *195*, 693–702. https://doi.org/10.1534/genetics.113.154138.

113. Garcia-Erill, G., and Albrechtsen, A. (2020). Evaluation of model fit of inferred admixture proportions. Mol. Ecol. Resour. *20*, 936–949. https://doi.org/10.1111/1755-0998.13171.

114. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A.R., Bender, D., Maller, J., Sklar, P., de Bakker, P.I.W., Daly, M.J., et al. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. *81*, 559–575. https://doi.org/10.1086/519795.

115. Liu, X., and Fu, Y.-X. (2020). Stairway Plot 2: demographic history inference with folded SNP frequency spectra. Genome Biol. *21*, 280. https://doi.org/10.1186/s13059-020-02196-9.

116. Excoffier, L., Dupanloup, I., Huerta-Sánchez, E., Sousa, V.C., and Foll, M. (2013). Robust demographic inference from genomic and SNP data. PLoS Genet. *9*, e1003905. https://doi.org/10.1371/journal.pgen.1003905.

117. Korunes, K.L., and Samuk, K. (2021). pixy: Unbiased estimation of nucleotide diversity and divergence in the presence of missing data. Mol. Ecol. Resour. *21*, 1359–1368. https://doi.org/10.1111/1755-0998.13326.

118. Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841–842. https://doi.org/10.1093/bioinformatics/btq033.

119. McLaren, W., Gil, L., Hunt, S.E., Riat, H.S., Ritchie, G.R.S., Thormann, A., Flicek, P., and Cunningham, F. (2016). The Ensembl Variant Effect Predictor. Genome Biol. *17*, 122. https://doi.org/10.1186/s13059-016-0974-4.

120. Cingolani, P., Platts, A., Wang, L.L., Coon, M., Nguyen, T., Wang, L., Land, S.J., Lu, X., and Ruden, D.M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. Fly (Austin) *6*, 80–92. https://doi.org/10.4161/fly.19695.

121. Haller, B.C., and Messer, P.W. (2023). SLiM 4: Multispecies Eco-Evolutionary Modeling. Am. Nat. *201*, E127–E139. https://doi.org/10.1086/723601.

122. McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., et al. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. *20*, 1297–1303. https://doi.org/10.1101/gr.107524.110.

123. Ersmark, E., Orlando, L., Sandoval-Castellanos, E., Barnes, I., Barnett, R., Stuart, A., Lister, A., and Dalén, L. (2015). Population Demography and Genetic Diversity in the Pleistocene Cave Lion. Open Quat. *1*, 1–15. https://doi.org/10.5334/oq.aa.

124. Dabney, J., Knapp, M., Glocke, I., Gansauge, M.-T., Weihmann, A., Nickel, B., Valdiosera, C., García, N., Pääbo, S., Arsuaga, J.-L., et al. (2013). Complete mitochondrial genome sequence of a Middle Pleistocene cave bear reconstructed from ultrashort DNA fragments. Proc. Natl. Acad. Sci. USA *110*, 15758–15763. https://doi.org/10.1073/pnas.1314445110.

125. Boessenkool, S., Hanghøj, K., Nistelberger, H.M., Der Sarkissian, C., Gondek, A.T., Orlando, L., Barrett, J.H., and Star, B. (2017). Combining bleach and mild predigestion improves ancient DNA recovery from

bones. Mol. Ecol. Resour. *17*, 742–751. https://doi.org/10.1111/1755-0998.12623.

126. Gilbert, M.T.P., Tomsho, L.P., Rendulic, S., Packard, M., Drautz, D.I., Sher, A., Tikhonov, A., Dalén, L., Kuznetsova, T., Kosintsev, P., et al. (2007). Whole-genome shotgun sequencing of mitochondria from ancient hair shafts. Science *317*, 1927–1930. https://doi.org/10.1126/science.1146971.

127. Carøe, C., Gopalakrishnan, S., Vinner, L., Mak, S.S.T., Sinding, M.H.S., Samaniego, J.A., Wales, N., Sicheritz-Pontén, T., and Gilbert, M.T.P. (2018). Single-tube library preparation for degraded DNA. Methods Ecol. Evol. *9*, 410–419. https://doi.org/10.1111/2041-210X.12871.

128. Fortes, G.G., and Paijmans, J.L.A. (2015). Analysis of Whole Mitogenomes from Ancient Samples. Methods Mol. Biol. *1347*, 179–195. https://doi.org/10.1007/978-1-4939-2990-0_13.

129. Mak, S.S.T., Gopalakrishnan, S., Carøe, C., Geng, C., Liu, S., Sinding, M.-H.S., Kuderna, L.F.K., Zhang, W., Fu, S., Vieira, F.G., et al. (2017). Comparative performance of the BGISEQ-500 vs Illumina HiSeq2500 sequencing platforms for palaeogenomic sequencing. GigaScience *6*, 1–13.

130. Liu, S., Westbury, M.V., Dussex, N., Mitchell, K.J., Sinding, M.S., Heintzman, P.D., Duchêne, D.A., Kapp, J.D., von Seth, J., Heiniger, H., et al. (2021). Ancient and modern genomes unravel the evolutionary history of the rhinoceros family. Cell *184*, 4874–4885.e16. https://doi.org/10.1016/j.cell.2021.07.032.

131. Smit, A.F.A., Hubley, R., and Green, P. (2013–2015). RepeatMasker Open-4.0. https://www.repeatmasker.org/.

132. Picard https://broadinstitute.github.io/picard/.

133. Pečnerová, P., Garcia-Erill, G., Liu, X., Nursyifa, C., Waples, R.K., Santander, C.G., Quinn, L., Frandsen, P., Meisner, J., Stæger, F.F., et al. (2021). High genetic diversity and low differentiation reflect the ecological versatility of the African leopard. Curr. Biol. *31*, 1862–1871.e5. https://doi.org/10.1016/j.cub.2021.01.064.

134. Meisner, J., and Albrechtsen, A. (2019). Testing for Hardy–Weinberg equilibrium in structured populations using genotype or low-depth next generation sequencing data. Mol. Ecol. Resour. *19*, 1144–1152. https://doi.org/10.1111/1755-0998.13019.

135. Orlando, L., Ginolhac, A., Zhang, G., Froese, D., Albrechtsen, A., Stiller, M., Schubert, M., Cappellini, E., Petersen, B., Moltke, I., et al. (2013). Recalibrating *Equus* evolution using the genome sequence of an early Middle Pleistocene horse. Nature *499*, 74–78. https://doi.org/10.1038/nature12323.

136. Achilli, A., Olivieri, A., Pellecchia, M., Uboldi, C., Colli, L., Al-Zahery, N., Accetturo, M., Pala, M., Hooshiar Kashani, B., Perego, U.A., et al. (2008). Mitochondrial genomes of extinct aurochs survive in domestic cattle. Curr. Biol. *18*, R157–R158. https://doi.org/10.1016/j.cub.2008.01.019.

137. Waples, R.K., Albrechtsen, A., and Moltke, I. (2019). Allele frequency-free inference of close familial relationships from genotypes or low-depth sequencing data. Mol. Ecol. *28*, 35–48. https://doi.org/10.1111/mec.14954.

138. Bhatia, G., Patterson, N., Sankararaman, S., and Price, A.L. (2013). Estimating and interpreting F_{ST}: the impact of rare variants. Genome Res. *23*, 1514–1521. https://doi.org/10.1101/gr.154831.113.

139. Zoonomia Consortium (2020). A comparative genomics multitool for scientific discovery and conservation. Nature *587*, 240–245. https://doi.org/10.1038/s41586-020-2876-6.

140. Chen, L., Qiu, Q., Jiang, Y., Wang, K., Lin, Z., Li, Z., Bibi, F., Yang, Y., Wang, J., Nie, W., et al. (2019). Large-scale ruminant genome sequencing provides insights into their evolution and distinct traits. Science *364*, eaav6202. https://doi.org/10.1126/science.aav6202.

141. Kumar, S., and Subramanian, S. (2002). Mutation Rates in Mammalian Genomes. Proc. Natl. Acad. Sci. USA *99*, 803–808.

142. Timmins, R.J., Hedges, S., and Robichaud, W. (2016). *Pseudoryx nghetinhensis*. The IUCN Red List of Threatened Species eT18597A46364962. file:///C:/Users/soft/Downloads/10.2305_IUCN.UK.2020-1.RLTS.T18597A166485696.en-1.pdf.

143. Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics *27*, 2987–2993. https://doi.org/10.1093/bioinformatics/btr509.

144. Rasmussen, M.S., Wiuf, C., and Albrechtsen, A. (2024). Inferring drift, genetic differentiation, and admixture graphs from low-depth sequencing data. Preprint at bioRxiv. https://doi.org/10.1101/2024.01.29.577762.

145. Busing, F.M.T.A., Meijer, E., and Leeden, R.V.D. (1999). Delete-m Jackknife for Unequal m. Stat. Comput. *9*, 3–8.

146. Kyriazis, C.C., Robinson, J.A., and Lohmueller, K.E. (2023). Using computational simulations to model deleterious variation and genetic load in natural populations. Am. Nat. *202*, 737–752. https://doi.org/10.1086/726736.

147. Pedersen, C.T., Lohmueller, K.E., Grarup, N., Bjerregaard, P., Hansen, T., Siegismund, H.R., Moltke, I., and Albrechtsen, A. (2017). The Effect of an Extreme and Prolonged Population Bottleneck on Patterns of Deleterious Variation: Insights from the Greenlandic Inuit. Genetics *205*, 787–801. https://doi.org/10.1534/genetics.116.193821.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Biological samples		
Saola samples	This study	See Table S5
Chemicals, peptides, and recombinant proteins		
AmpliTaq Gold DNA Polymerase	Applied Biosystems	Cat.#10055114
Critical commercial assays		
DNeasy extraction kit	QIAGEN	Cat.# 69504
MinElute PCR Purification Kit	QIAGEN	Cat.# 28004
EB Buffer	QIAGEN	Cat.# 19086
QIAquick columns	QIAGEN	Cat.# 28115
Qubit dsDNA HS Assay	Life technologies	Cat.# Q33230
PfuTurbo DNA Polymerase	Agilent	Cat.# 600250
AMPure XP beads	Beckman Coulter	Cat.# 10136224
Deposited data		
Saola raw sequence data and assembly	This study	SRA BioProject ID PRJNA688353
Saola mitochondrial genome	Genbank	GenBank: NC_020616.1
Cow sequencing data	Das et al.[83]	SRA Sample ID SAMN04201346
Bos taurus genome	Genbank	GenBank: GCA_000003055.5
Capra hircus genome	NCBI RefSeq	RefSeq: GCF_001704415.1
Bubalis bubalis genome	NCBI RefSeq	RefSeq: GCF_000471725.1
Tragelaphus imberbis genome	Genbank	GenBank: GCA_006410775.1
Balaenoptera acutorostrata genome	NCBI RefSeq	RefSeq: GCF_000493695.1
Software and algorithms		
COPE v1.2.4	Liu et al.[84]	https://github.com/dhlbh/COPE
QUAKE v0.3	Kelley et al.[85]	https://github.com/davek44/Quake
FastUniq v1.1	Xu et al.[86]	https://sourceforge.net/projects/fastuniq/
Platanus v1.2.4	Kajitani et al.[35]	http://platanus.bio.titech.ac.jp/platanus-assembler
RACA	Kim et al.[38]	https://github.com/ma-compbio/RACA
UCSC Toolkit	Kuhn et al.[87]	http://genome.ucsc.edu
BUSCO v3.0.1	Simão et al.[36]	https://busco.ezlab.org
RepeatMasker v4.0.5	Smit et al.[88]	https://www.repeatmasker.org
Augustus v2.5.5	Stanke and Morgenstern[89]	https://github.com/Gaius-Augustus/Augustus/tree/master
glimmerHMM v3.0.3	Majoros et al.[90]	https://ccb.jhu.edu/software/glimmerhmm/
EVidenceModeler v1.1.1	Haas et al.[91]	https://github.com/EVidenceModeler/EVidenceModeler
LAST v984	Kiełbasa et al.[92]	https://github.com/lpryszcz/last
MULTIZ v10.6	Blanchette et al.[93]	https://www.bx.psu.edu/miller_lab/
RaxML v8.2.9	Stamatakis[94]	https://cme.h-its.org/exelixis/web/software/raxml/
ASTRAL-III v5.5.6	Mirarab et al.[95]	https://github.com/smirarab/ASTRAL
PALEOMIX rev. 6c44fa53	Schubert et al.[96]	https://github.com/MikkelSchubert/paleomix
AdapterRemoval v2.3.2	Schubert et al.[97]	https://github.com/MikkelSchubert/adapterremoval

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
BWA v0.7.17	Li and Durbin[98]	https://github.com/lh3/bwa
Picard tools v2.24	Broad Institute	https://broadinstitute.github.io/picard/
Samtools v1.11.0	Li et al.[99]	https://samtools.github.io
Genmap v1.2.0	Pockrandt et al.[100]	https://github.com/cpockrandt/genmap
SATC	Nursyifa et al.[101]	https://github.com/popgenDK/SATC
ANGSD v0.933-102-g7d57642	Korneliussen et al.[102]	https://github.com/ANGSD/angsd
PCAngsd v0.98	Meisner and Albrechtsen[103]	https://github.com/Rosemeis/pcangsd
mapDamage v2.2.1	Jónsson et al.[104]	https://ginolhac.github.io/mapDamage/
MAFFT v7.407	Katoh and Standley[105]	https://mafft.cbrc.jp/alignment/software/
ModelTest-NG v0.1.7	Darriba et al.[106]	https://github.com/ddarriba/modeltest
BEAST v2.7	Bouckaert et al.[107]	https://www.beast2.org
Tracer v1.7.1	Rambaut et al.[108]	http://beast.community/tracer
Bcftools 1.13	Danecek et al.[109]	https://samtools.github.io/bcftools/
Winsfs v0.7.0	Rasmussen et al.[110]	https://github.com/malthesr/winsfs
NGSrelate v2	Hanghøj et al.[111]	https://github.com/ANGSD/NgsRelate
NGSadmix	Skotte et al.[112]	http://www.popgen.dk/software/index.php/NgsAdmix
evalAdmix v0.962	Garcia-Erill and Albrechtsen[113]	https://github.com/GenisGE/evalAdmix
PLINK v.1.9	Purcell et al.[114]	https://www.cog-genomics.org/plink/
PSMC v.0.6.5-r67	Li and Durbin[46]	https://github.com/lh3/psmc
stairwayplot v2	Liu and Fu[115]	https://github.com/xiaoming-liu/stairway-plot-v2
fastsimcoal2 v2.7.0.93	Excoffier et al.[116]	https://cmpg.unibe.ch/software/fastsimcoal27/
pixy 1.2.7.beta	Korunes and Samuk[117]	https://pixy.readthedocs.io/en/latest/
bedtools v2.29.2	Quinlan and Hall[118]	https://bedtools.readthedocs.io/en/latest/
VEP v108.2	McLaren et al.[119]	https://www.ensembl.org/info/docs/tools/vep/index.html
snpEff v.4.3	Cingolani et al.[120]	https://pcingola.github.io/SnpEff//
SLiM 4.0.1	Haller and Messer[121]	https://messerlab.org/slim/
GATK v.4.1.7	McKenna et al.[122]	https://gatk.broadinstitute.org/hc/en-us

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

The sample material was primarily composed of samples that were collected during the initial discovery of the saola and that were used for its formal description and preliminary genetic analyses in Van Dung et al.[1] These were subsequently stored at the University of Copenhagen, Denmark, and supplemented with extra samples collected by the Central Institute for Natural Resources and Environmental Studies (CRES) in Vietnam and subsequently stored there. The material was predominantly small bits of degraded skin, hair or bone sampled from trophies in indigenous hunter villages, plus a small number of tissues, suspended in an unknown preservation buffer at -80 °C. See Table S4 for an overview of the sample material.

METHOD DETAILS

DNA extraction and sequencing library preparation

DNA was isolated from buffer-preserved soft tissue samples using the Qiagen DNeasy extraction kit (Qiagen, Hilden, Germany, [QIAGEN]) following the manufacturer's guidelines. Briefly after initial maceration of the tissue, a ca. 25 mg sample was digested with Buffer ATL and Proteinase K overnight at 56 °C using a thermomixer set to 800 rpm. After pelleting any undigested material, the supernatant was added to Buffer AL/ethanol as recommended, transferred to a DNeasy Mini spin column and centrifuged at 8,000 rpm. Next, DNA was purified using Buffer AW1 and Buffer AW2. Finally, DNA was incubated in the membrane with elution Buffer AE for 5 min at room temperature and centrifuged it at 8,000 rpm for 1 minute.

The bone samples were handled separately from the rest of samples in the ancient DNA facilities at the Globe Institute, University of Copenhagen. The bone pieces were manually fragmented with a hammer, then placed in 2 mL microcentrifuge tubes and briefly

washed with 1 mL of 10% dilution of commercial bleach solution, followed by several rinses with molecular biology grade water to remove residual bleach. The bone fragments were subsequently pre-digested for 1 hour at 56 °C with rotation, in 1 mL of freshly made digestion buffer containing 0.5 M EDTA (pH 8), 1 M urea and 10 μg/μl proteinase K solution.[123] This first lysate, which is typically rich in microbial DNA, was discarded, after which a second digestion with rotation at 56 °C was performed overnight in 1 mL of the same digestion buffer. After incubation, the tubes were allowed to equilibrate to room temperature, then centrifuged on a bench-top centrifuge at 6,000 x g for 2 minutes. The lysate supernatant was purified using the MinElute PCR Purification Kit (Qiagen), with modifications to retain short DNA fragments based on.[124] Final elution with EB buffer (Qiagen) was performed in two steps, with an incubation period of 10 minutes at 37 °C before each centrifugation, for a final volume of 35 μl.

The dried tissue samples were processed under strict clean laboratory conditions at the Globe Institute, University of Copenhagen. Tissue samples were placed into 2 mL microcentrifuge tubes and washed with a 3.5% bleach solution, ethanol and ddH2O, following.[125] The material was processed following[126] DNA extraction protocol. After washing, tissue samples were pre-digested for 15 minutes at 37 °C with rotation, in 1 mL of freshly made digestion buffer containing 10mM Tris-HCl (pH 8.0), 10mM Sodium chloride, 5mM Calcium chloride, 2.5mM EDTA (pH 8.0), 2% w/v SDS, 1nM DTT and 10 μg/μL of proteinase K. The pre-digestion lysate was discarded. A second digestion with 1 mL of the same buffer and settings was performed overnight. After incubation, additional treatment with phenol chloroform was performed following.[127] The supernatant was then purified using MinElute column with modified PB buffer and eluted using 2 washes in pre-warmed 18 μL buffer EB (QIAGEN) - with 3 min of incubation time at 37°C.[124] The DNA concentration of each extract was verified on a Qubit (ng/μL).

The integrity of all DNA extracts was initially visualized on a Bioanalyzer. This revealed that sample 9264 contained surprisingly high quality DNA, with observable DNA fragments of up to ca 15 kb in size (Figure S1A), leading us to decide to generate a de novo reference genome for that sample, based around Illumina libraries constructed with insert sizes spanning a range from 250 bp to 5 kb, and sequenced using the Illumina HiSeq X10 platform. For short insert size libraries (< 1 Kb), DNA was fragmented to the desired size (250 bp and 500 bp), and then end-repaired and ligated to the Illumina pair-end adaptors. Further procedures including size selection, purification and PCR amplification were conducted accordingly to get the final libraries. For the mate-paired libraries (2 kb and 5 kb), the fragmented DNA of target length were circularized and the remaining linear DNA were digested to better enrich the targeted fragments. After that, the circularized fragments were fragmented again and biotinylated to prepare for the following adaptor ligation steps.

Illumina sequencing libraries were constructed for all remaining samples for quality assessment (specifically endogenous DNA content estimation), via low coverage sequencing using Illumina platforms. If necessary, DNA was first fragmented using a Bioruptor (Diagenode) after which Illumina libraries were constructed following.[128] Specifically fragments were end repaired with T4 Ploynuclease kinase, T4 Polymerase, after which a blunt end adapter was ligated using 0.5 μM of Illumina P5 and P7 adapter mix (20 μM each). Adapter-ligated libraries were then purified with MinElute (Qiagen, Hilden, DE) and polymerase chain reaction (PCR) amplified with Primer IS4 and individual index primers in a total volume of 15 μL using AmpliTaq Gold polymerase (Applied Biosystems). Finally, the indexed libraries were purified using MinElute columns and eluted in 20 μl elution buffer EB. Multiple libraries were combined together into 3 pools, normalized to 10 nM, and sequenced across 3 lanes of Illumina HiSeq 2500 using either SR100 chemistry or Illumina HiSeq X10 using PE 150 chemistry.

Following initial screening we elected to generate more data from a subset of the samples. In this case BGISeq compatible libraries were prepared following,[129] which consists of three reactions: end repair of DNA fragments by means of T4 DNA Polymerase and T4 Polynucleotide Kinase; ligation of adapters with T4 DNA ligase, and fill-in of adapters with Bst Warmstart Polymerase. Enzymes are heat-inactivated between steps. Finalised libraries were purified on QiaQuick columns (Qiagen, Hilden, Germany). Final elution was performed in two steps with an incubation period of 10 minutes at 37 °C before each centrifugation step, for a total volume of 40 μl. Libraries were given 18 cycles of amplification in a total volume of 100 μL, containing 10 U AmpliTaq Gold polymerase (Applied Biosystems, Foster City, CA), 1× AmpliTaq Gold buffer, 2.5 mM $MgCl_2$, 0.4 mg/mL bovine serum albumin (BSA), 0.2 mM each dNTP, 0.2 μM BGI common primer, 0.2 μM indexed reverse primer, and 20 μL of DNA library template. Following amplification, libraries were purified according to manufacturer's instructions using either a QIAquick PCR Purification kit (Qiagen, Hilden, Germany), or Agencourt AMPure XP beads (Beckman Coulter) with a bead-to-DNA ratio of 1.8. DNA was eluted in a final volume of 35 μL EB buffer. During silica column purification, the column was incubated for 10 minutes at 37°C prior to centrifugation. DNA concentrations were measured with a Qubit fluorometer using a dsDNA high sensitivity assay (Life Technologies, Carlsbad, CA) and a TapeStation 2200 (Agilent, Santa Clara, CA). Amplified libraries corresponding to the most poorly preserved saola samples were subjected to an extra step of size selection to remove high molecular weight fragments of DNA that were likely to be exogenous contaminant sequences. In order to retain only fragments between 100 and 800 bp, the libraries were treated twice with Agencourt AMPure XP beads (Beckman Coulter), first with a bead-to-DNA ratio of 0.6 to discard fragments above 800 bp, then the supernatant of that treatment was cleaned with a 1.8 bead-to-DNA ratio to remove fragments below 100 bp. The libraries were then sequenced on BGISeq-500 platforms using PE 100 chemistry.

Following our initial data analyses we also elected to improve the total genomic depth of coverage for 3 samples from the Southern population. Illumina libraries were built using 8-20 μl of extract in a final reaction volume of 50 μl following the Santa Cruz Single Stranded protocol.[130] This protocol consists in opening the double stranded structure into single stranded using Single Stranded Binding proteins during a 95 °C step followed by adapter ligation using T4 DNA Ligase and T4 Polynucleotide Kinase. Libraries were then purified on MinElute columns and eluted to a final volume of 42 μL. Libraries were then amplified in three replicates with 20 cycles in a volume of 50 μL. Each replicate contained 2.5 U PfuTurbo DNA Polymerase (Agilent, Santa Clara, CA), 0.5 μM

forward index, 0.5 μM indexed reverse primer, 10x PFU buffer, 25 μM dNTP, BSA (Bovine Serum Albumin), and 5 μL of sample. Following amplification, replicates were combined and purified using MinElute columns and eluted in a final volume of 40 μL. DNA concentrations were measured with a Qubit fluorometer using a dsDNA high sensitivity assay and fragment size profile was visualized using Bioanalyzer. Finally, amplified libraries were sequenced in a NovoSeq Illumina platform using PE 150 chemistry.

QUANTIFICATION AND STATISTICAL ANALYSIS

Reference genome assembly and annotation

Sequence reads from the different insert size libraries underwent different data pretreatment regimes. For the libraries with 250 bp inserts, we used COPE v.1.2.4[84] to overlap and merge the read pairs, resulting in ≈95% of reads being overlapped. We discarded the reads that could not be overlapped. We then used QUAKE[85] v.0.3 to error correct the reads of the other 6 libraries based on the 17-mer frequency from the reads of the overlapped 250 bp libraries. We used FASTUNIQ v1.1[86] to identify and remove duplicate reads. Subsequently, we used Platanus v1.2.4[35] to assemble the genome, by first building contigs using the 250 and 500 bp insert libraries, to which the reads mate-pair are mapped to, removing reads that mapped with too short insert sizes for the library expectation. Finally, all reads were used for scaffolding requiring at least 3 reads support to link contigs. Post initial assembly, we used method RACA[38] that takes advantage of both paired-end read mapping and comparative genome information from related genomes to create chromosome-scale assemblies, to further generate Predicted Chromosome Fragments (PCFs). We aligned the assembled saola genome and the *Capra hircus* genome (RefSeq: GCF_001704415.1) to the *Bos taurus* genome (GenBank: GCA_000003055.5) using LAST v984.[92] We then chained and netted the alignments to add synteny information and to form high quality blocks using the UCSC toolkit.[87] After that, we aligned the mate-pair reads to the saola genome, and used both the synteny and read pair information to connect the scaffolds into chromosomal fragments with a minimum block size of 100 kb. We finally used BUSCO v. 3.0.1[36] to evaluate genome completeness, with the Laurasiatheria lineage database (orthoDB v9).

Transposable elements were identified and masked using RepeatMasker v4.0.5.[131] Then, homolog prediction was conducted with amino acid data sets available for sheep (*Ovis aries*), cow (*Bos taurus*) and humans, while Augustus v2.5.5[89] and glimmerHMM v3.0.3[90] were used to do de novo gene prediction. Subsequently EVidenceModeler v1.1.1[91] was used to merge the results of the different methods to obtain the final genome annotation.

Phylogenetic placement of the saola

We aligned the saola genome and other genomes including *Bubalus bubalis* (water buffalo, RefSeq: GCF_000471725.1), *Tragelaphus imberbis* (lesser kudu, GenBank: GCA_006410775.1), *Capra hircus* (domestic goat, RefSeq: GCF_001704415.1), and *Balaenoptera acutorostrata* (minke whale, RefSeq: GCF_000493695.1) to the available reference sequences of *Bos taurus* (cattle, GenBank: GCA_000003055.5) using LAST v984[92] and combined them together into a multiple genome alignment using MULTIZ v10.6.[93] We conducted a sliding window-based phylogenetic analysis with a window size of 100 kb and step length of 50 kb using the cattle genome as reference, using the mafsInRegions utility from the UCSC toolkit[87] to define and extract the regions from the multiple alignment. We used the GTRCATI model in RaxML v8.2.9[94] to estimate the phylogenetic tree for each window, and inferred the final species tree from the phylogenies of each window using ASTRAL-III v5.5.6.[95]

Mapping of resequenced samples

We demultiplexed the samples with a custom perl script, excluding all reads whose index sequence did not exactly match any of the expected index, since these reads are more likely to also contain sequencing errors elsewhere. We identified adapters using AdapterRemoval v2.3.2,[97] and then mapped the samples to the RACA-scaffolded saola reference genome using PALEOMIX rev. 6c44fa53,[96] trimming for adapters/low quality bases and merged using AdapterRemoval, and mapped using BWA v0.7.17[98] using the backtrack algorithm, and post-processing using Picard tools v2.24[132] and Samtools v1.11.0.[99] PALEOMIX was run using default parameters, except in that AdapterRemoval adapters were overridden based on the observed adapter sequences, that read merging was performed using the "conservative" algorithm, and that unmapped reads and detected PCR duplicates were flagged but not removed during mapping.

The resulting BAMs were subsequently filtered to eliminate unmapped reads, secondary and supplementary alignments, QC failed reads, and PCR duplicates. In addition, for paired end reads that had not been merged, read pairs in improper orientations or mapping to two contigs, with inferred insert sizes less than 190 bp or greater than 500bp, and reads with fewer than 35 bp or 70% of their length mapped were eliminated. In cases where one read in a pair was filtered, the entire pair was discarded.

Reference genome masking for population genomic analyses

For the population-level analyses we used a combination of different approaches to identify and exclude problematic regions in the reference genome that are more likely to contain mapping or genotyping errors.[133]

Reference genome-based masking

First, we excluded all PCFs shorter than 1Mb. We furthermore masked all transposable elements and other repetitive regions identified with RepeatMasker v4.0.5. Finally, we used genmap v. 1.2.0[100] to infer mappability across the genome, assuming reads of length 70 and allowing up to 2 mismatches per read. We excluded all sites with mappability lower than 1.

Mapped data-based masking

We identified sex-associated PCFs based on the normalised depth of the samples mapped using SATC,[101] and confirmed all of them had been predicted to be part of the X chromosome. We excluded sex-associated PCFs from all analyses.

We then estimated combined depth per site using ANGSD v. 0.933-102-g7d57642,[102] separately for a group of samples with average depth below 5, and another group of samples with average depth above 5. We excluded all sites with depth below ½ of the median or above 3/2 of the median for either of the two groups of samples.

Finally, we excluded regions exhibiting an excess of heterozygosity. We identified SNPs with strong excess of heterozygosity using the per-site Hardy-Weinberg equilibrium (HWE) test implemented in PCAngsd v 0.98[103,134] which corrects for population structure. As input, we used genotype likelihoods for a preliminary set of SNPs inferred with ANGSD using the GATK model (-gl 2),[122] excluding bases with calling quality below 30 (-minQ 30), reads with mapping quality below 25 (-minMapQ 25), and keeping only sites with minor allele frequency above 0.05 (-minmaf 0.05) and with p-value for being a SNP above 1e-6 (-SNP_pval 1e-6) and excluding SNPs resulting from a transition mutation (-rmTrnas 1). We then applied the population structure-aware HWE test in PCAngsd with default values, with the minimum average partial implemented in PCAngsd identifying 1 as the optimal number of principal components to model the population structure. We selected sites with extreme and significant deviation of HWE in the direction of excess of heterozygosity, as those with an inbreeding coefficient $F < -0.9$ and with a minimum p-value in the likelihood-ratio test of 1e-6, and excluded from further analysis regions within 10 kb of these SNPs.

Ancestral state estimation in the saola reference genome

We obtained the publicly available sequencing data for a Holstein cattle sample (SRA Sample ID SAMN04201346;[83]) and mapped it to the saola reference genome, using the same mapping pipeline and post-mapping filtering as previously described for the saola samples. We then used ANGSD to produce a sequence in fasta format, using the most common base at each position (-dofasta 2) after excluding reads with mapping quality below 25 (-minmapq 25) and bases with base quality below 30 (-minq 30).

DNA damage patterns and error rate estimation

We used mapDamage v2.2.1[104] to estimate patterns of DNA damage in each sample from the mapped data, downsampling to 100,000 reads.

We estimated error rates across the genome after site filtering using the 'perfect individual' approach[135] implemented in ANGSD.[102] Error rates are estimated as excess distance of the samples to an outgroup, relative to the distance from a sample's consensus sequence that is assumed to be error free and the outgroup. In this case, we used 'perfect individual' the saola reference genome, and used the cattle consensus sequence (see previous subsection) as the outgroup.

Mitochondrial genome reconstruction

We determined the consensus sequences (the most common bases for each site) of the mitochondrial genome for each individual. First, the mitochondrial genome of the *de novo* genome sequenced individual (NVq1) was obtained by mapping a subset of its raw reads (sampled to ca. 10X coverage) to the pre-existing saola mitochondrial genome available in Genbank (NC_020616.1),[29] and using ANGSD to get the consensus fasta sequence as that with the most bases supporting it (-dofasta 2), trimming 5 bp from both ends of each read and requiring a minimum depth per site of 5.[102] This mitochondrial genome was included in the reference genome to which all samples were mapped. We then obtained the consensus sequence for each of the samples similarly, using ANGSD to call the consensus with most reads supporting it. We excluded reads with mapping quality below 25 and bases with base calling quality below 30, and sites where more than 5% of the reads supported a different base or sites with depth below 30.

Mitochondrial tree

The consensus sequences were aligned with published mitogenomes available from GenBank representing a cow (NC006853), an impala (NC020675), a lesser kudu (NC020619), a goat (NC005044), a water buffalo (NC049568), and a Cape buffalo (NC020617). The sequences were aligned in MAFFT v7.407.[105] The alignment of 17,109 sites was used in ModelTest-NG v0.1.7[106] to determine the best-fitting substitution scheme and GTR+I+G4 was selected based on the Bayesian Information Criterion. We reconstructed a phylogenetic tree in BEAST v2.7[107] using strict molecular clock with a per generation rate of 2.043E-8,[136] the Coalescent Bayesian Skyline tree model, a chain length of 10,000,000 samples, storing every 5,000 samples and pre-burnin of 30. We verified the MCMC convergence in Tracer v1.7.1,[108] with all but one Effective Sample Size values above 200. We generated a maximum clade credibility tree in TreeAnnotator based on the Common Ancestor heights, a 30% burn-in and a 0.7 posterior probability limit. For plotting, only the cow was used as an outgroup.

Population SNP calling and genotype likelihood estimation

We called SNPs only in the sites that passed the filters described above (see "reference genome masking for population genomic analyses"), and furthermore excluded all transition mutations to be able to include samples with damaged DNA and maximise the sample size. We used ANGSD to call SNPs and estimate genotype likelihoods using the GATK model (-gl 2[122]), excluding bases with base calling quality below 30 (-minQ 30), reads with mapping quality below 25 (-minMapQ 25), and keeping only sites with minor allele frequency above 0.05 (-minmaf 0.05) and with *p*-value for being a SNP above 1e-6 (-SNP_pval 1e-6) and excluding SNPs

resulting from a transition mutation (-rmTrnas 1), using the genotype likelihoods to infer the major and minor alleles (-domajorminor 1). This resulted in a total of 644,546 transversion SNPs called for 26 saola.

Genotype calling

We called genotypes in 8 samples for which the average depth after filtering was > 6X and the estimated error rate in transversion mutations was lower than 0.001. We used bcftools 1.13[109] and called genotypes on all sites, and subsequently removed sites based on the above described filters (see "reference genome masking for population genomic analyses"). We furthermore removed indels and multi allelic SNPs, and set to missing all heterozygous calls where either of the two alleles had less than 3 reads supporting it. Further filtering of transition mutations and masking of low depth calls with varying threshold was done differently depending on certain analyses, and is specified in the corresponding section.

Duplicates, relatedness and inbreeding inference

We used two complementary methods to estimate pairwise relatedness between samples. We first used ibsRelate,[137] an allele-frequency free estimator of relatedness based on three statistics that are functions of the 2DSFS between pairs of samples. We performed an initial estimation of these statistics using all sites, by first estimating per sample site allele frequencies using ANGSD,[102] without calling major and minor allele (-dosaf 1), using the GATK genotype likelihood model (-gl 2), and excluding bases with base calling quality below 30 (-minQ 30) and reads with mapping quality below 25 (-minMapQ 25), and excluding transition mutations (-noTrans 1). We subsequently used winsfs v0.7.0[110] to estimate pairwise 2DSFS between samples and calculate the three ibsRelate statistics. Furthermore, we also applied ibsRelate as implemented in NGSrelate v2[111] with a preliminary data set where the duplicate samples identified with the winsfs based estimates had been merged. For this, we first estimated genotype likelihoods separately for each population with ANGSD, keeping only common SNPs (-snp_pvalue 1e-6 and -minmaf 0.05), excluding bases with base calling quality below 30 (-minQ 30) and reads with mapping quality below 25 (-minMapQ 25), and excluding transition mutations (-rmTrans 1). We then used the genotype likelihoods as input for NGSrelate v2. Using only common SNPs allowed to reduce the impact of sequencing errors are allowed to identify a further pair of duplicate samples that had not been captured in the winsfs-based estimates (Figure S2A).

After identifying duplicate samples and excluding or merging them (Figure S2A; Table S4), we used NGSrelate v2[111] to jointly estimate relatedness and inbreeding separately within each of the two populations. We again estimated genotype likelihoods for variable sites separately within each of the two populations using ANGSD,[102] and with the same filtering as previously described, and used the estimated genotype likelihoods as input for NGSrelate v2.

Inference of population structure

We first estimated the covariance matrix from the genotype likelihoods of the above described dataset with 644,546 transversion SNPs for 26 saola samples using PCAngsd.[103] We used 1 principal component in the PCAngsd iterative algorithm to estimate population allele frequencies (-e 1), which was detected as optimal by the built-in Velicer's minimum average partial test from PCAngsd.

We furthermore used NGSadmix[112] to estimate admixture proportions from genotype likelihoods for the same 644,546 SNPs and 26 samples. We assumed a value of K of 2, 3 and 4 and, and for each K, we did multiple independent runs until either convergence was reached, by obtaining 3 runs within 2 likelihood units of the maximum likelihood run, or 100 runs were done without convergence. For the runs that converged, we assessed the model fit of the admixture proportions by estimating the pairwise correlation of residuals between individuals with evalAdmix v0.962.[113]

Estimation of heterozygosity

We used several approaches and filtering schemes to estimate heterozygosity and validate the estimates. We first estimated heterozygosity from genotype likelihoods, by first estimating a saf file with ANGSD, without calling major and minor allele (-dosaf 1), using the GATK genotype likelihood model (-gl 2), and excluding bases with base calling quality below 30 (-minQ 30) and reads with mapping quality below 25 (-minMapQ 25). We did this twice, first using all mutations and another excluding transition mutations (-noTrans 1). We then estimated the individual site frequency spectrum (SFS) with winsfs.[110] To assess the quality of the estimates, we visualized jointly the heterozygosity estimates with the estimated error rates and average depth of each sample (Figure S4B). Based on visual inspection of this, we considered the heterozygosity estimates for samples with sequencing depth higher than 6 and transversion error rates in transversion mutations lower than 0.001.

Furthermore, for the samples for which we had called genotypes, we estimated heterozygosities from the genotype calls. We did this using a bcf file containing both variable and fixed sites, and excluded sites with a sequencing depth below 10 and heterozygous calls with less than 3 reads supporting either of the two alleles. We again repeated the estimation with and without transitions. We then assessed the concordance in the estimates between the two methods, and furthermore for the samples with good enough quality to obtain reliable estimates when using all mutations, we confirmed they fit the expected ratio of all mutations to transversions of 3/1 (Figure S4A). Based on this and with the aim to maximize the number of samples used, we chose to present results obtained excluding transition mutations but rescaled by multiplying them by 3, so they reflect the average heterozygosity when considering all mutations.

We obtained a compilation of genome-wide heterozygosities estimated for different animals to compare with the saola's estimates (Table S3 in Liu et al.[130]).

Inference of runs of homozygosity

We estimated ROHs for the 8 good quality samples for which we had called genotypes. We used as input a genotype file with transversion mutations variable within the 8 samples, setting as missing sites where the sample has depth below 6 and heterozygous sites with less than 3 reads supporting either read masked, with a total of 588,703 sites kept. To account for the high variability in missingness across our samples, we split each sample in an individual binary PLINK file and kept only sites where that sample had data. We only called ROH in scaffolds with enough contiguity to identify ROHs, by removing scaffolds shorter than 10 Mb, and consequently adapted the total genome size when calculating the proportion of genome within ROHs. The number of sites retained for each sample ranged between 261,807 and 571,961. We used PLINK v.1.9[114] to call ROHs, using windows of 100 SNPs (–homozyg-window-snp 100) required at least 1 SNP per 100 kb to call a ROH (–homozyg-density 100), allowing at most 1 heterozygous SNP per window within a ROH (–homozyg-window-het 1), extending ROH to adjacent homozygous variants (–homozyg 'extend') with a minimum length of 1 Mb to call a ROH (–homozyg-kb 1000) and using otherwise default settings. These settings were chosen after comparing and validating the results of different settings by visualizing genomic plots of called ROH, SNP heterozygosity, SNP density and finding these to be within the most reliable (Figure S4C).

Population site frequency spectrum estimation

We estimated site allele frequency (.saf) files for selected subsets of samples, with differing filtering criteria depending on the analyses they would be used in. We first estimated saf files for 8 samples in the Northern population with estimated error rates in transversion mutations lower than 0.0005, and after excluding a sample inferred to be first degree relative of other included samples (NVq9, see Figure S4D). We estimated saf with ANGSD (-dosaf 1), excluding bases with base call quality below 30 (-minq 30) and reads with mapping quality below 25 (-minmapq 25), excluding transition mutations (-noTrans 1) and using the GATK genotype likelihood model (-gl 2). We then used winsfs,[110] using 500 blocks per window (-w 500) and otherwise default settings, to estimate the SFS for the Northern population, and used it as input for estimating the saola's recent effective population sizes trajectories in stairwayplot v2 (see below). We did not estimate an SFS for the Southern population with this filtering criteria, because it would result in only 3 samples kept, which is a too low sample size to obtain any reliable information on recent population size trajectories.

We then estimated a saf file for the Southern and Northern populations, with 3 and 5 individuals respectively, after removing any samples with depth below 7 or error rates in transition mutations higher than 0.001. We estimated a saf file for each of the populations using the same settings as described above, and used winsfs, also with the same settings as described above, to estimate the 2DSFS between the Northern and Southern population. This 2DSFS was used to estimate F_{ST} and to estimate the joint demographic history of the populations with fastsimcoal2 (see below).

F_{ST} estimation

We quantified the genetic differentiation between the two populations using the Hudson's estimator of the fixation index F_{ST},[138] as implemented in winsfs. We used two different approaches to obtain the population 2DSFS used for the estimation of SFS. First, we used the individual 2DSFS between the two highest depth samples from each population, NVq1 and SHu1, obtained from the called genotypes for all mutation types, filtering sites where either of the two samples had depth below 10 or sites where any of the two samples had an heterozygous call with less than 3 reads supporting either of the two alleles. The second approach was using the 2DSFS estimated from the saf file using multiple samples per population, as previously described.

Comparison of saola genetic diversity with other species

Genome-wide heterozygosity

We obtained a pre-compiled dataset of genome-wide heterozygosities estimates for 321 animal species.[130] All of these quantify the same parameter: the proportion of nucleotide positions that are heterozygous for biallelic SNPs across a diploid genome. However, because they come from different studies, differences in filtering and software used for genotype calling can potentially have a significant impact on the heterozygosity estimate. For this reason, we re-estimated genome-wide heterozygosity for the saola sample NVq1 using two alternative pipelines that are used for a large proportion of the species in the compiled dataset. We refer to the pipelines as the "zoonomia pipeline", which is used in all samples from Zoonomia Consortium,[139] and the "Liu et al. pipeline", used for all samples from Liu et al.[130] and Zoonomia Consortium[140] Sample NVq1 was chosen because it is the same sample used to generate the saola reference genome which the resequencing data is mapped to, as is the case for the other studies we compared pipelines with, hence removing any possible reference bias.

In brief, for the "zoonomia pipeline" we used GATK v.4.1.7[122] HaplotypeCaller to call genotypes, using genotype banding at 0, 10, 20, 30, 40, 50 and 99 qualities and otherwise default settings. We then filtered to keep only SNP variants and invariant positions, and kept only sites with quality > 15. For the "Liu et al. pipeline", we used ANGSD v. 0.933-102 do-saf function to generate saf files, filtering for minimum base calling quality of 20, minimum read mapping quality of 20, a minimum and maximum depth of respectively $^1/_3$ and 2 times the average depth for the individual, and keeping only uniquely mapped reads.

Functional genetic diversity

We used the variant effect predictor (VEP) v108.2 software[119] to predict the functional impact of the saola's called variants, using the gene annotation that we generated as previously described. We used the genotype calls with a minimum depth of 10 and required at least 3 reads supporting either of the two alleles to keep a heterozygous call. We furthermore kept only the 3 samples with average

depth above 15X and error rates across all mutations lower than 0.002. We then counted for each sample the number of identified heterozygous variants in three categories predicted by their impact on the protein: for loss of function (LoF) mutations are those expected to have a high impact, disrupting the function, missense mutations are those that modify the coded amino acid and could impact the protein function, and silent mutations are those that do not change the amino acid sequence and are thus likely to not have a phenotypic impact. We also obtained a publicly available compilation of heterozygous sites across multiple samples for different mammal species for the same three categories (LoF, missense and silent sites; Table S4 in Liu et al.[130]), as a comparison with the inferred values in the saola. For comparability with this mammal species compilation, we redid the analyses for the saola annotating variants using snpEff v.4.3,[120] using the saola annotation to build a custom database and otherwise using default settings in snpEff. As a measure of the amount of segregating functional variation, we used the ratio of the count of LoF and missense heterozygous variants to the count of silent heterozygous variants within each individual.

Genetic diversity in windows across the genome

We estimated genetic diversity π in non-overlapping windows across the genome for scaffolds longer than 10 Mb using pixy 1.2.7.beta.[117] We used the jointly called genotype for six samples, three from the Northern population and three from the Southern population. We used all samples with error rates in transversion mutations lower than 0.001 and depth higher than 3 in the Southern population (sample IDs SHu1, STg2 and STg3), and selected the three samples with highest depth and error rates in transversion mutations below 0.001 in the Northern population (sample IDs NVq1, NVq2 and NVq3), to have a matched number of samples in both populations. We used genotype calls for both variable and non-variable sites, excluded transition mutations, set to missing genotypes with depth lower than 6 and heterozygous calls with less than 3 reads supporting either of the two alleles, and subsequently removed sites where 50 % of the samples or more had missing data. We then estimated π across windows for each population and for the combined population with all 6 samples, using different window sizes (see Figure 4A). We subsequently excluded windows that had less than the 0.05 quantile of non-missing sites for the corresponding window size.

Sharing of regions in ROHs across samples

We generated per sample bed files with the regions identified to be in ROH (see above), and used bedtools v2.29.2[118] multiinter to find overlaps between ROH of all samples keeping track of the population of origin of each. For each region with two or more overlapping ROHs, we inferred whether all samples in ROH had the same haplotype in the region. We considered only regions with at least 10 called variable sites, and for all pairs of samples in ROH defined they had the same haplotype if 98% of nonmissing genotypes in the region were the same between the pair. We did this three times, by considering only samples from each of the two populations, and considering samples from all populations.

Furthermore, we used bedtools to intersect the regions with different numbers of samples in ROH with the coding sequences from the gene annotation in gff format. For each category defined by the number of samples in ROH, we calculated the average overlap with coding sequence across all regions in that category. We again did this three times, by considering only samples from each of the two populations, and considering samples from both populations combined. Finally, to estimate the significance of deviations from genome-wide average of coding sequence in the different categories, we generated a null distribution by using bedtools shuffle for each individual bed file of regions in ROH, multi intersecting all individual's shuffled ROHs and finding the overlap with coding sequence of each ROH category. We did this 1,000 times to generate a null distribution, and for each ROH category and each sample grouping we calculated a 2-sided empirical *p*-value as the minimum of the fraction of shuffles with coding sequence percentage above or below the observed value times 2:

$$p = 2\,min\left(\frac{\sum_{i}^{1000} 1_{\{s_i > obs\}}}{1000}, \frac{\sum_{i}^{1000} 1_{\{s_i < obs\}}}{1000}\right),$$

where obs indicates the observed average overlap with coding sequence for the category, and s_i the average overlap for the category in the *i*th shuffled iteration.

Demographic history analyses

Assumptions of mutation rate and generation time

We used a mutation rate of 1.2e-8 mutations/generation[141] and a generation time of 6 years[142] to scale the estimates of demographic history. For SFS-based analyses where the SFS had been estimated excluding transition mutations, we used a rescaled mutation rate with the expected 1:3 ratio to 4e-9 mutations/generation.

Pairwise sequentially Markovian coalescent (PSMC)

We used PSMC v.0.6.5-r67[46] to infer long-term population size trajectories for the three samples with high depth and overall low damage patterns and error rates. We called genotypes individually for each sample with bcftools v1.15 consensus caller,[143] excluding reads with mapping quality below 25 and bases with base quality below 30. We subsequently excluded sites where the sample had less than $^1/_3$ of its average depth, or less than 6X if $^1/_3$ of the average depth was lower, sites where the depth was

more than twice the average depth, and heterozygous calls with less than 3 reads supporting either allele. We subsequently generated psmc input files and ran psmc with default settings.

Estimation of relative cross-coalescence rates using unphased inbred genomes with PSMC

We exploited the high proportion of the saola genomes in ROH to estimate population cross-coalescence rates. We identified genomic regions where each of the two high quality samples from the Northern population (NVq1 and NVq2) and the high quality sample from the Southern population (SHu1) were both in ROH longer than 1 Mb, which was around 120 Mb for each pair. We then used the genotype calls for all sites, including non-variable sites, to obtain the haplotype for each sample on these regions. We excluded any sites within these ROH where samples had an heterozygous call, that in a ROH region are majoritarily due to genotyping errors, or a minority of very recent de novo mutations that have no information on past coalescence rates. We then combined these in two pairs of inter-population pseudodiploid sequences, and used PSMC to estimate coalescence rates between the sequence, obtaining thus estimates of population cross coalescence rates through time.

We then used average within population coalescence rates as a denominator for calculating relative cross-coalescence rates through time. The reduced genomic region used and potential systematic differences between genomic regions in ROH and not in ROH made estimates of coalescence rates with the previous approach not comparable with those obtained from the whole genome (Figure 5A vs Figure S7B). We thus selected within each of the three samples regions of similar size and ascertainment criteria as the regions for which to estimate within population coalescence rates. For each sample we generated two regions, where each of the other two samples were in ROH longer than 1 Mb while the focal sample was not in ROH, in this case using a minimum ROH length of 0.5 Mb to maximise the amount of information of the distant past contained in the sequence. We thus generated two regions for each of the three samples, using regions in ROH in either of the other two samples. Finally, we subsampled the resulting regions until keeping a total amount of sequence of the same size as the length of the ROH intersection for that pair of samples. We extracted genotypes on these regions, from the same genotype calls as above, and ran PSMC with the same parameters.

Finally, we estimated relative cross coalescence rates by dividing the average coalescence rates between populations by the average coalescence rate within populations for each time segment.

Stairway plot

We used the folded SFS estimated using 8 samples from the Northern population (see above) to estimate the effective population size trajectories of the Northern population with stairwayplot v2,[115] using recommended settings.

Fastsimcoal

The divergence time between saola southern population and northern population was investigated using a coalescent simulation based method implemented in fastsimcoal2 v2.7.0.93.[116] To minimize potential bias arising when determining ancestral allelic states, we used the folded 2dSFS, based on the whole genome 2DSFS estimated with winsfs as previously described. We assumed a simple model only considering divergence time without gene flow between the two populations. For this model we ran 100 independent Fastsimcoal runs to find the best-fitting parameters yielding the highest likelihood, with 500,000 coalescent simulations per likelihood estimation (-n500000), 100 conditional maximization algorithm cycles (-L100), and minimum 100 observed SFS entry count taken into account in likelihood computation (-C100). A mutation rate of 4e-9 per site per generation and a generation time of 6 years were used to convert model estimates from coalescence units to years. Moreover, we used a non-parametric jackknife approach to estimate the uncertainty in the fitted model parameters. We used winsfs to estimate SFS in contiguous blocks of the genome of approximately the same length.[144] We then produced a total of 50 leave-one-out SFS splits by summing the estimated SFS of all but one block each. For each of the 50 splits we fitted again the demographic model with fastsimcoal v2.7.0.93, using the maximum likelihood parameters inferred from the whole genome as initial guesses for the optimization. For each split we fitted 10 independent runs, and selected the run with the maximum likelihood of the final parameters. Finally, we used the maximum likelihood parameters of each split to estimate standard errors using the block jackknife estimator for unequal size.[145]

Simulations of genetic load

We used SLiM 4.0.1[121] to perform individual-based forward simulations of the saola demographic history with a non-Wright Fisher implementation. To avoid assumptions on the unknown saola's life-history traits, we used a simple model with non-overlapping generations where in each generation N offspring are created by crossing randomly selected males and females in the population, where N is the population size as specified by the saola demography. Subsequently all parental individuals die, while offspring survive to reproduce in the next population with probability proportional to their absolute fitness. We used the demographic history inferred with stairwayplot v2.0 for the Northern population (Figure 5B). Because low sample sizes prevented us from inferring a reliable recent demographic history for the Southern population, we used the demography from the Northern population rescaled to $^{2}/_{3}$. This was based on the fact that PSMC suggests very similar trajectories for both populations, with the Southern having approximately $^{2}/_{3}$ lower population sizes. We used the time of the first decrease in population size 3,334 generations ago as the split time between the two populations. We also simulated a counterfactual demography, where saola population declines have started much more recently. In this case we followed the same initial demographic trajectories and split times, but upon the split the two populations were kept at a constant population size for longer and only started decreasing exponentially 10 generations before present (Figure S9C).

We simulated only coding regions and, based on the observed gene content and size in the saola genome, simulating 18,442 genes with a length of 1,935 bp, organized in chromosomes mimicking the saola's genome PCF. We set a recombination rate of 1e-3

between genes in the same chromosome, and free recombination between chromosomes. In total, we simulated 35,703,710 bp of coding sequence.

We used a mutation rate of 1.2e-8 per generation, and a ratio of 2.31 deleterious mutations to 1 neutral mutation.[146] For the deleterious mutations, we duplicated all simulations with two previously proposed models for similar simulation set-ups. The first one, described and used in Pérez-Pereira et al.,[48] samples deleterious selection coefficients (s) from a gamma distribution with mean 0.2 and shape parameter 0.33. Dominance coefficients (h) are then sampled from a uniform distribution in the range from 0 to e^{-ks}, where k is a parameter set such that the average h is 0.283 (Figure S8A, left). This model, therefore, enforces a negative relationship between s and h, such that the more deleterious mutations are more recessive, but allows the less deleterious mutations to be both recessive and additive. The second model is described in Kyriazis et al.[146] and samples from a gamma distribution with mean 0.0131 and shape 0.186, and is further augmented with a 0.3 % of lethal mutation (s = 1). Then h is totally determined for a given s such that h = 0 for s > 0.1, h = 0.05 for 0.1 > s >= 0.01, h=0.2 for 0.01 > s >= 0.001 and h = 0.45 for s < 0.001 (Figure S8A, right). The models therefore differ in the average deleteriousness of mutations, with the Pérez-Pereira model assuming more deleterious mutations, and in the relationship between s and h. While both agree with the broadly observed negative relationship between h and s, the Pérez-Pereira model allows for more variable h such that less deleterious mutations can potentially be recessive or partially recessive, in opposition to the Kyriazis model. Given the influence the marginal and joint distribution of s and h can potentially have in the dynamics of genetic load, and the outstanding uncertainty in which models better fit that observed in natural populations, we chose to present results with these two different models to evaluate how sensitive results are to the assumed distributions.

Genetic load was calculated per generation as genetic load and realized load following the definitions in Bertorelle et al.[49] The realized genetic load captures the loss in fitness of an individual due to the burden of deleterious mutations, and is given by

$$realized\ load = \sum_{i \in HET} s_i h_i + \sum_{j \in HOM} s_j$$

where HOM is the set of genotypes homozygous for the derived alleles and HET is the set of heterozygous genotypes. The selection coefficients are s and h, such that h codes as 0, 0.5, 1 when recessive, additive and dominant respectively. Genetic load measures the total amount and magnitude of deleterious variations an individual carries without considering its dominance coefficient, although it can also be defined as the realized load assuming additivity of all mutations, and for this reason it is sometimes called 'additive genetic load'[147]

$$genetic\ load = \sum_{i \in HET} s_i 0.5 + \sum_{j \in HOM} s_j$$

Note that the genetic load does not depend on demographic changes by themselves as it is not affected by changes in genome-wide heterozygosity, but only depends on the number of derived alleles. Therefore, if selection is not removing deleterious variants then the genetic load will not be affected, while if there is genetic purging then the genetic load will be reduced. We used this to develop a measure of purging, in which we normalize the genetic load at each generation (x) by the genetic load at a certain reference time point.

$$purging = \frac{genetic\ load_{\,t = x}}{genetic\ load_{t = ref}}$$

By using as reference the generation before the decrease in population sizes, we can measure the relative change in genetic load due to demographic changes. A relative decrease in genetic load will indicate that there is purging.

We also simulated different captive conservation management to assess how accumulated genetic load through the population decline would influence the probability of success of different scenarios. For this, we used the end-points of the previous simulations (the present time) as the starting point. We simulated different captive conservation management to assess how accumulated genetic load would influence the probability of success of different scenarios. For each scenario we varied the number of founding individuals (4, 12 or 24, having in all cases half of the founder male and half female) and the population of origin of the founding individuals (all from the Northern population, all from the Southern population, and a mixed breeding program where half individuals come from the Northern population and half from the Southern). This results in 9 different scenarios tested. For each scenario, the reproduction at generation *t* happens by generating 2 N_t offspring from randomly sampled pairs of males and females, where N_t is the number of individuals in the scenario in generation *t*. Starting thus from the number of founders, the population sizes will increase or decrease driven by the absolute fitness of the individuals and random fluctuations. We stopped the simulations of a scenario if any of the following conditions were met: i) all individuals or either of all males or females individuals died (in that case the scenario outcome was labeled as "Extinct"), ii) when the population size increased over 1,000 individuals, or iii) when 100 generations had passed without either of these two conditions happening (in either two cases the scenario outcome was labeled as "Survival"). We performed 200 independent simulations, from which we calculated the average proportion of simulations with survival and extinction for each scenario.

CellPress
OPEN ACCESS

Supplemental figures

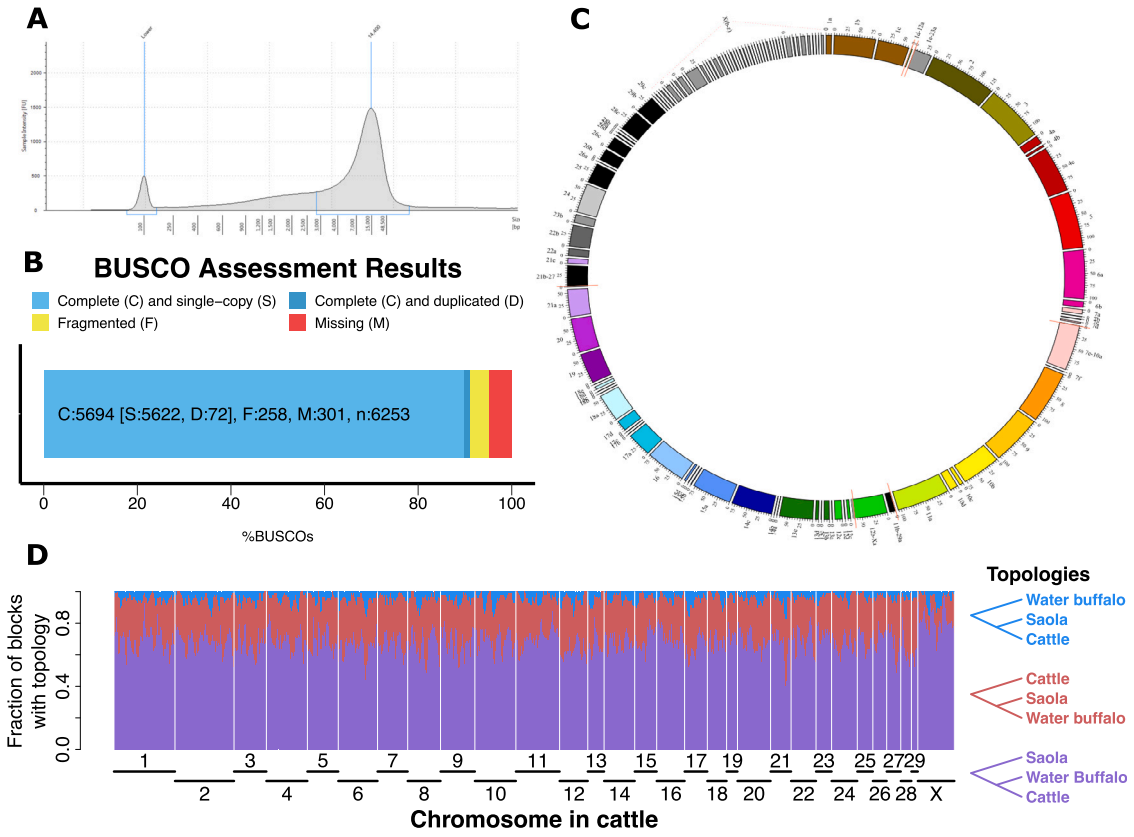


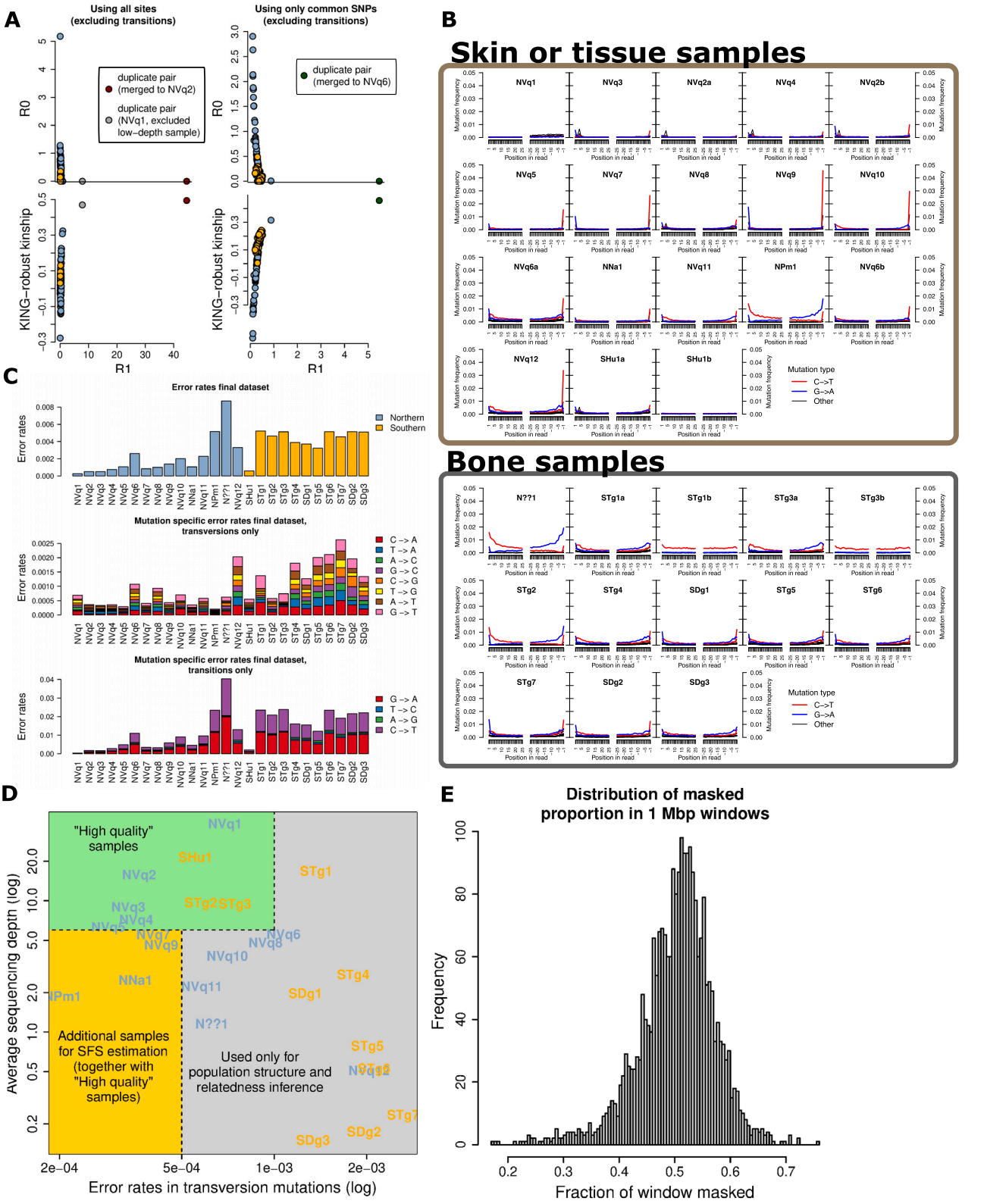
Figure S1. Saola reference genome assembly and phylogenetic variation along the genome, related to Figure 1 and STAR Methods

(A) Distribution of DNA fragment lengths for the sample NVq1/9264 that was used to construct the saola reference genome.

(B) Summary of BUSCO results for assessing the completeness of the saola genome. The lineage dataset is laurasiatheria_odb9 (creation date: 2016-02-13, number of species: 25).

(C) PCFs constructed by RACA. The circle displays the mapping between saola PCFs and cattle chromosomes, which are represented by distinct colors. The PCFs were named according to their mapped cattle chromosome.

(D) Proportion of windows supporting each of the three main topologies, differing in the placement of the saola lineage relative to the water buffalo and cattle, in the tree used to find the saola phylogenetic placement (Figure 1). Proportions were calculated in sliding windows of 5 Mb with a step size of 500 kb.



(legend on next page)

Figure S2. Quality control of the resequencing dataset, related to Figure 2 and STAR Methods

(A) Inference of duplicate samples using the allele frequency-free method ibsRelate. The expected values for duplicates are R0 = 0, R1 > 1, and KING-robust = 0.5. The left plot shows the analyses based on SFS estimated from winSFS. The right plot shows the analyses based on estimates with NGSRelate V2 from genotype likelihoods SNPs with MAF > 0.05, which is less sensitive to sequencing errors.

(B) Map damage plot for different samples to explore the presence of deamination in the ends of DNA fragments, indicative of DNA damage. Samples were sequenced in two batches, and samples that were duplicated and thus merged to a single individual are indicated by the suffixes "a" and "b" (see STAR Methods and Table S5). Panels are grouped by the material of the sample (only sample NVq1 comes from tissue, and it is grouped together with the rest of skin samples).

(C) Relative error rates, estimated as the excess distance of the sample's reads to the consensus allele in a cow sample (as a proxy to the ancestral state) relative to the distance of the consensus sequence of the highest depth sample (NVq1; the "perfect" individual) and the inferred ancestral state. Error rates thus reflect both sequencing errors and DNA damage. The top panels show overall error rates across all mutations, while the middle and bottom panel show mutation-specific error rates for transversion and transition mutations, respectively.

(D) Visualization of the sample subsets used in different analyses and the criteria for delimiting the sample subsets based on average sequencing depth and error rate in transition mutations. Note that sample NVq9 was removed from the SFS estimation due to being a close relative of sample NVq3.

(E) Distribution of masked proportion of 1 Mbp genomic windows after applying the sites filtering criteria described in Table S6.

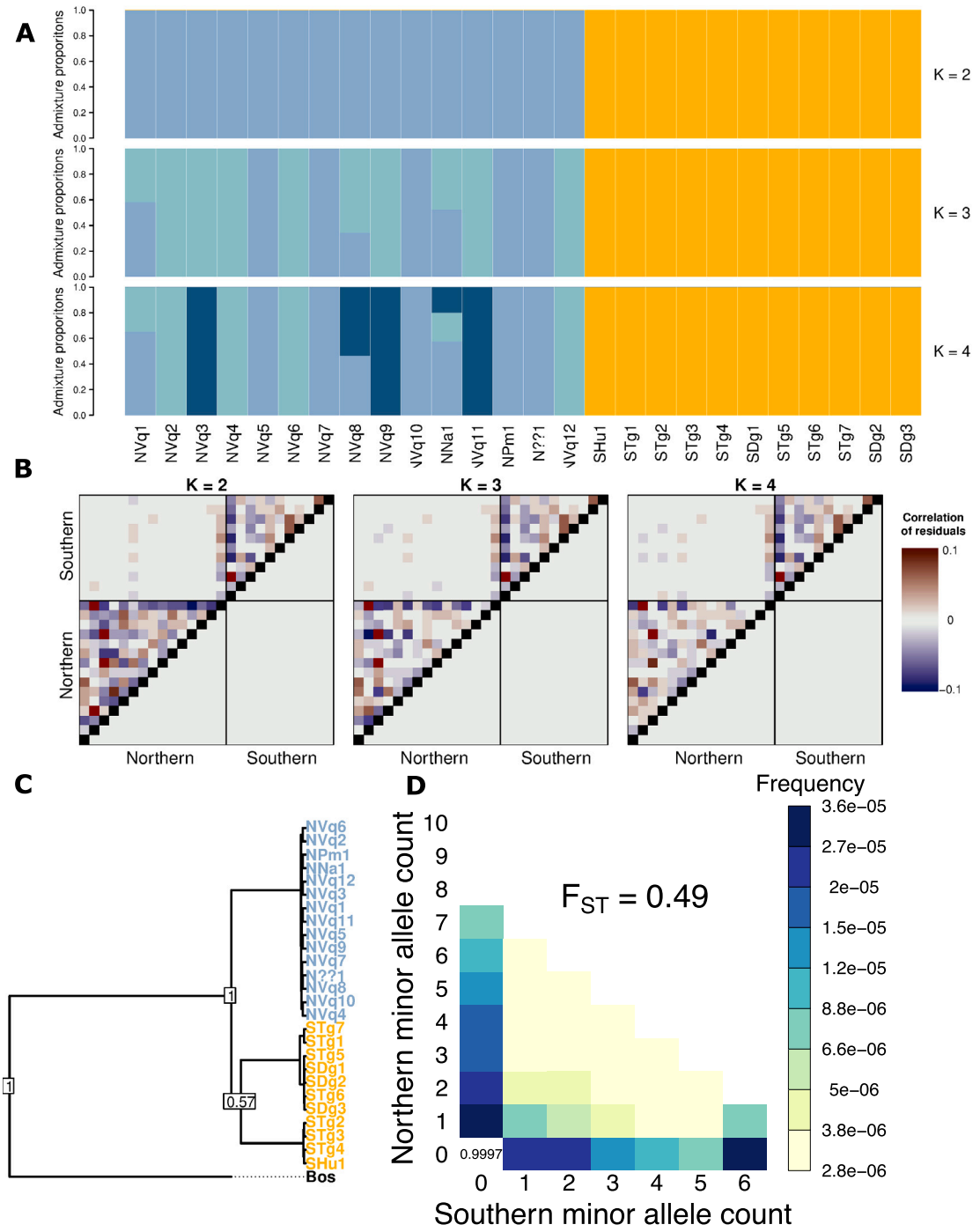


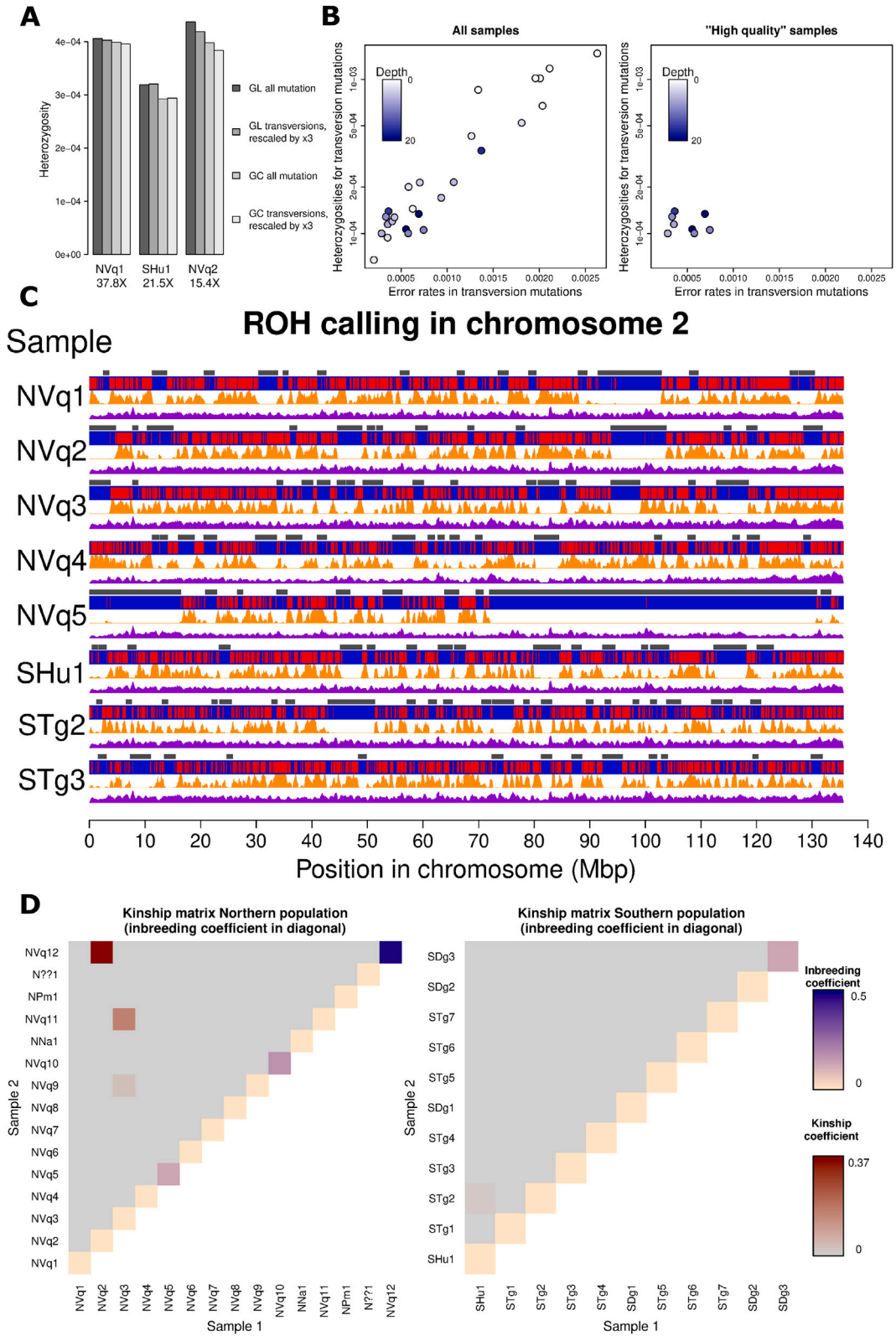
Figure S3. Saola population structure, related to Figure 2

(A) NGSadmix for K = 2 to K = 4 for the 26 saolas in the final dataset. Values of K larger than 2 detect substructure within the populations that likely reflect distant and close relatives rather than population substructure.

(B) Evaluation of the admixture proportions for K = 2 to K = 4 shown in (A), as the correlation of residuals estimated with evalAdmix. Non-zero correlations of residuals are indicative of a bad model fit. Assuming K = 2, there is no systematic non-zero correlation within populations, indicating the major population structure is captured at K = 2. Some correlation between individuals remains due to the presence of close relatives and inbred individuals, and potentially population substructure that we cannot resolve with the current sample size.

(C) mtDNA tree, with samples colored by population based on the genetic clustering of the whole-genome analysis. Internal nodes are labeled with the posterior probability as node support.

(D) Two-dimensional SFS (2DSFS) between the northern and southern populations, using the genotype calls for high-quality samples and excluding transition mutations. The high F_{ST} estimated from the SFS (printed on the upper diagonal), is reflecting the low amount of shared variation between the two populations.



(legend on next page)

Figure S4. Saola genetic diversity, inbreeding, and relatedness, related to Figure 2

(A) Comparison of heterozygosity estimation with different approaches, using samples with sequencing depth >15 and error rates lower than 0.002. GL is the heterozygosity estimate based on genotype likelihoods with winsfs, and GC is based on genotype calls, using only sites with depth above 10 and with at least 3 alleles supporting each allele in heterozygous calls. "All mutations" indicate that all mutation types are used for the heterozygosity estimation, while "transversions" indicate that transition mutations are excluded, and the estimated heterozygosity is rescaled by the expected ratio of 1:3.

(B) Impact of error rates and sequencing depth on heterozygosity estimates when considering all 26 samples (left) and using only the selected subset of 8 high-quality samples (right). The high-quality samples are defined as those having error rates in transversion mutations lower than 0.001 and average sequencing depth above 6×. When considering all samples, there is a clear linear relationship between heterozygosities and error rates, but this relationship disappears when considering only good-quality samples. The heterozygosity estimates rescaled to reflect the expected value using all mutations for good-quality samples are shown in Figure 2F.

(C) Visualization of ROH calling using the largest chromosome as an example for the eight high-quality samples. For each sample from top to bottom, the dark gray segments show the regions inferred to be in ROH > 1 Mbp, the middle panel shows in blue all homozygous variants in the chromosome with heterozygous variants overlapped in red, and the bottom panels show the density of heterozygous calls (orange) and the SNP density (purple) in 10 kb windows across the chromosome.

(D) Estimation of relatedness (in upper triangle cells) and inbreeding (in diagonal cells) jointly estimated within each population with NGSrelateV2 with the maximum likelihood allele frequency-based method.

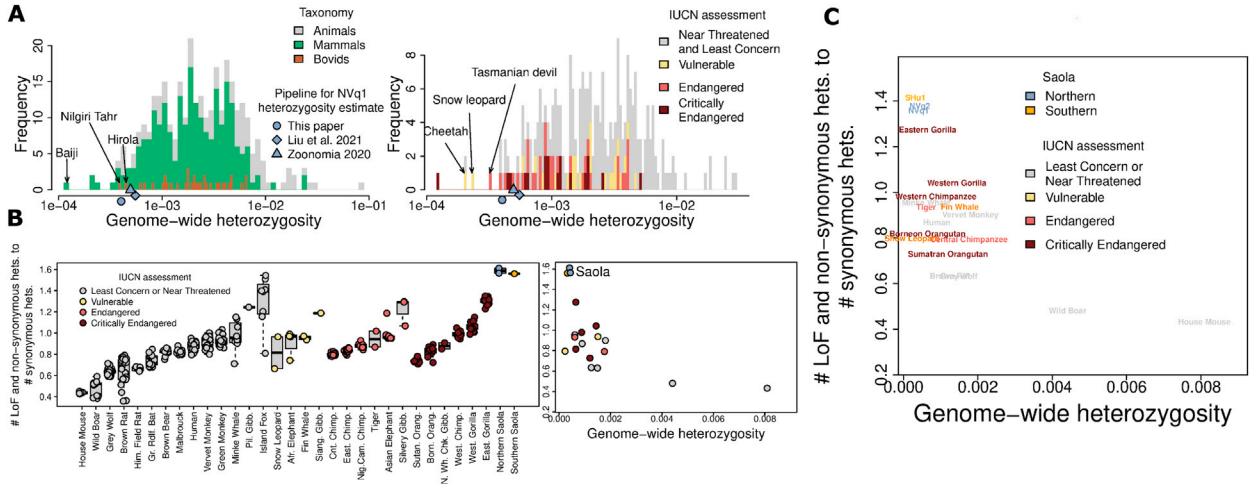


Figure S5. Comparison of saola genome-wide and functional genetic diversity with other species, related to Figure 3

(A) Same as Figure 3A, but showing for saola three estimates of genome-wide heterozygosity for sample NVq1 using different pipelines that are used in the compiled multispecies dataset.

(B) Same plot as Figure 3B, but the estimates for saola are based on variant annotation done with snpEff instead of VEP.

(C) Same plot as in Figure 3C, but with the name of the species instead of a point. It shows the ratio of number of potentially deleterious heterozygote sites (LOF and non-synomous) to synonymous variants plotted estimate of genome-wide heterozygosity. Species colored by conservation status.

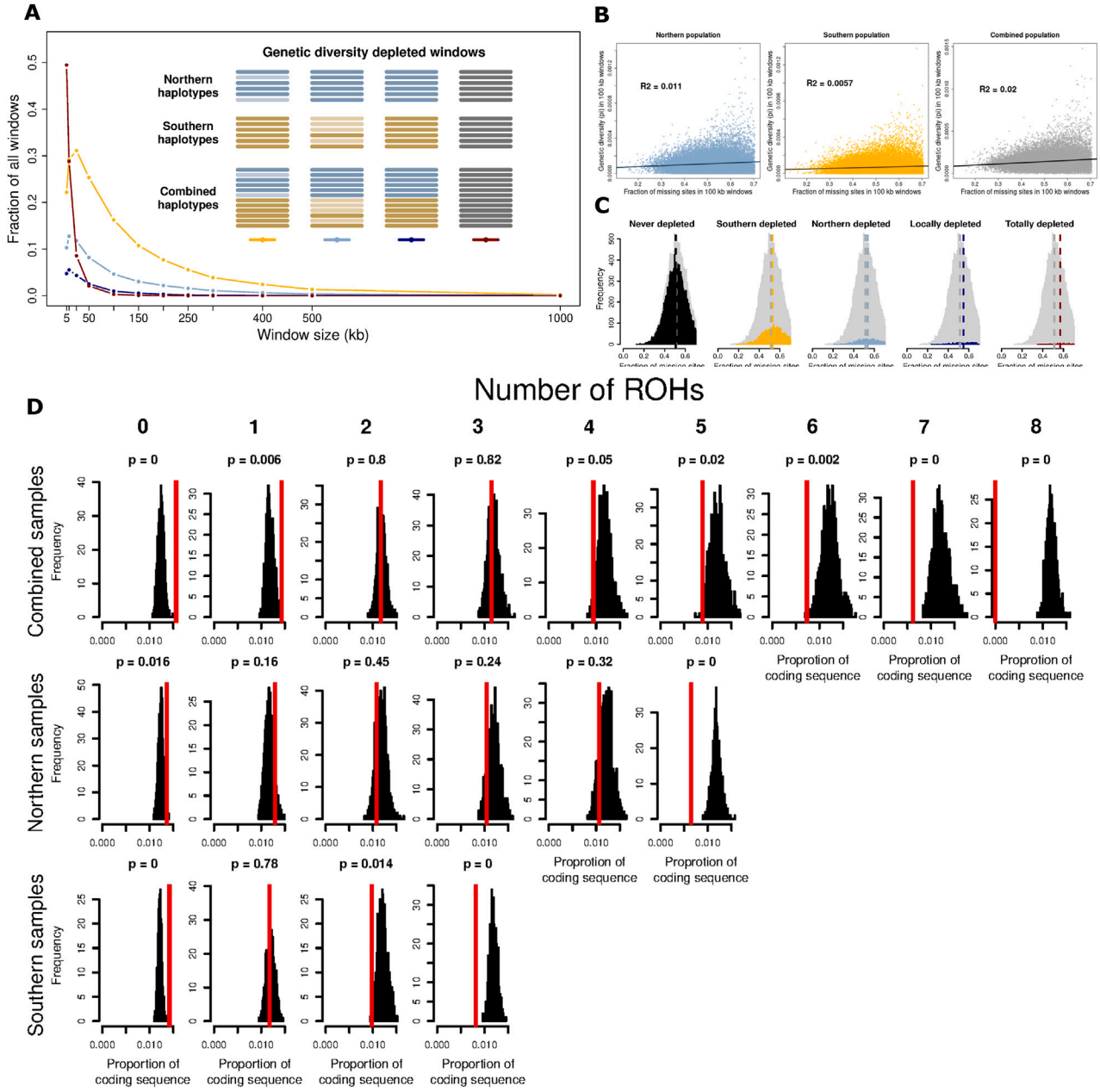


Figure S6. Robustness and significance of genomic diversity loss along saola genomes, related to Figure 4

(A) Fraction of windows without any genetic diversity only in the southern population (orange), only in the northern population (light blue), in both populations but not in a combined population (dark blue), and in a combined population (dark red), across different window sizes. The legend depicts a cartoon of how the classification is done based on haplotypes, where differences in color tone and brightness represent different haplotypes. The results with a window size of 100 kb correspond to what is shown in Figure 4C.

(B) Correlation between the fraction of missing sites and diversity for the analysis of 100 kb windows shown in Figure 4A. There is a consistent but weak positive correlation.

(C) Distribution of missingness for the 4 categories of depleted windows that are shown as an upset plot in Figure 4B. The left panel ("never depleted") corresponds to windows not fitting any of the 4 categories, i.e., windows where there is some genetic diversity within both populations. "Locally depleted" corresponds to windows where neither population has genetic diversity, but there is diversity in the combined population, while "totally depleted" corresponds to windows where there is no diversity in either population nor in the combined population. The background gray distribution shows the distribution across all windows. The dotted lines indicate the mean missingness in each category. There is a bias where windows with more missingness are more likely to be depleted, but the bias is stronger in the totally depleted category. Thus, our conclusion that sites depleted of genetic diversity are usually not shared between populations is robust to this bias, since it tends to increase the frequency of windows depleted in both categories in a higher proportion than it increases the rest.

(legend continued on next page)

(D) Permutation test of significance on the differences in overlap of ROH and coding sequence shown in Figure 4D. The test is done by shuffling the location of ROH for each individual and for each permutation, calculating the overlap between the shuffled ROHs to obtain the categories of the number of ROHs, for which the average proportion of coding sequence is then estimated. This was repeated 1,000 times to obtain the distribution shown in the histograms. Each histogram shows the overlap with coding sequence for regions with a certain number of shared ROH (grouped in columns) and a certain grouping of populations (grouped in the rows). The red vertical lines indicate the observed values for each category. The *p* values indicated above each histogram are calculated with a two-sided test (see STAR Methods).

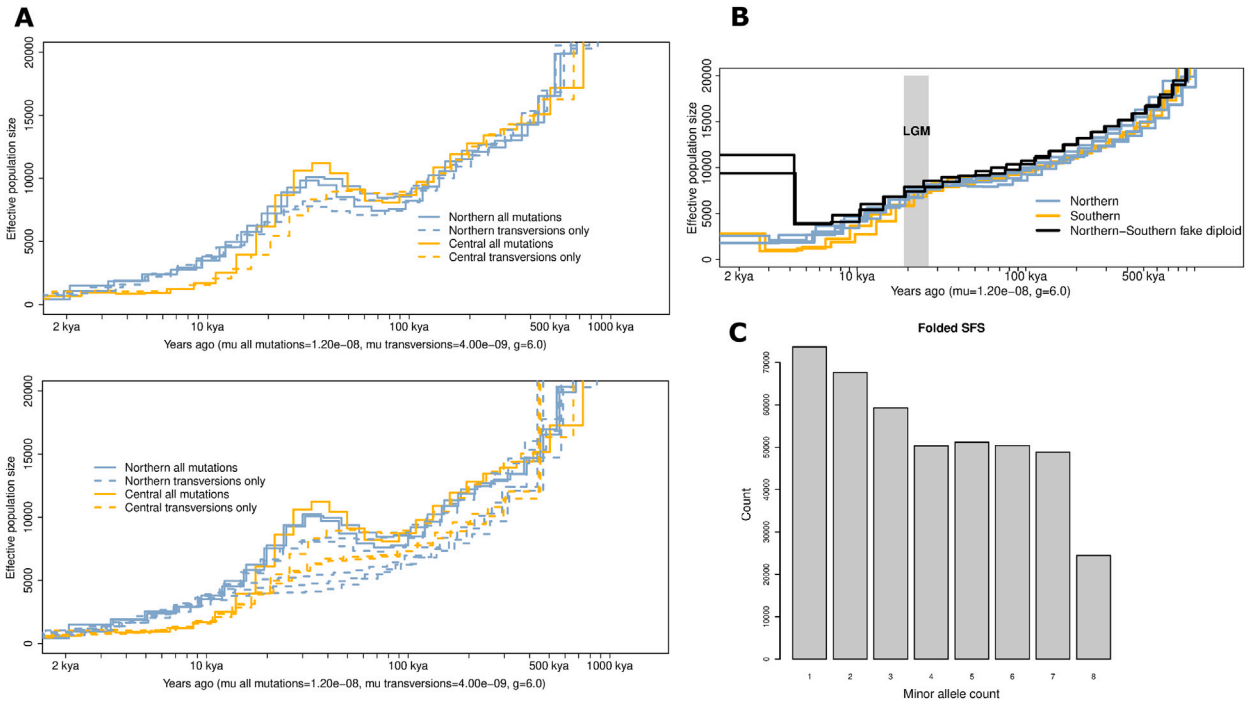


Figure S7. Supplemental demographic history of saola, related to Figure 5

(A) PSMC using all mutations and using only transversion mutations. The upper panel shows only the 3 high-depth samples with low enough DNA damage to run PSMC with all mutations for comparison of the trajectories. The lower panel shows all samples for which genotypes were called in the version without transition mutations (transversions only).

(B) PSMC estimates used for the cross-coalescence rates estimation shown in Figure 5A. Effective population sizes are inversely proportional to coalescence rates by a constant scaling factor, and the relative cross-coalescence is thus calculated as the average of the synthetic "northern-southern diploid" estimates divided by the average of the "northern" and "southern" estimates for each discrete time period.

(C) Folded SFS for the northern population, used as input for the stairway plot analyses shown in Figure 5B, estimated excluding transition mutations for those samples from the northern population with estimated error rates in transversion mutations lower than 0.001 and excluding two first degree relatives. Only variable sites are shown.

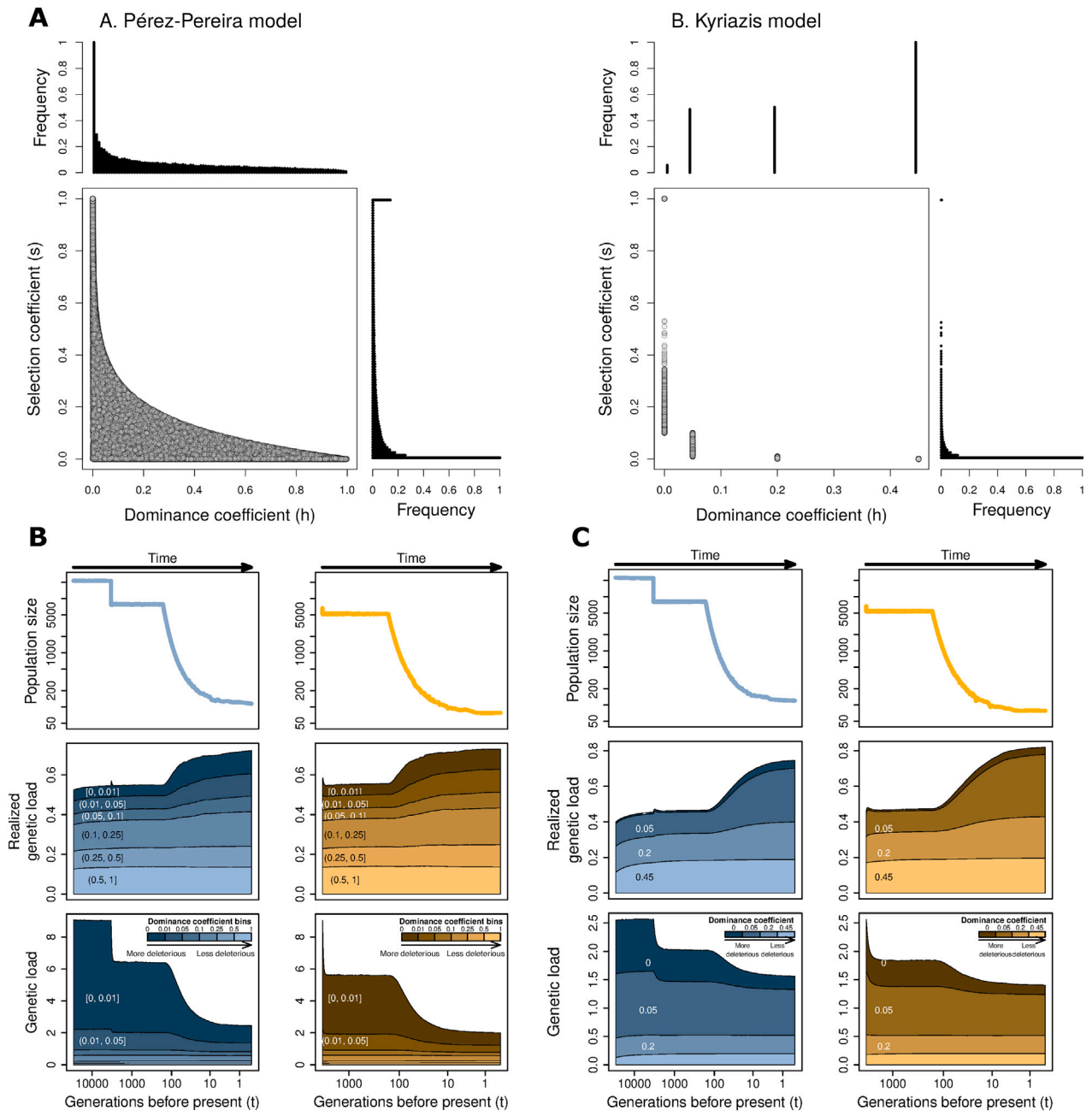


Figure S8. Models for the joint distribution selection and dominance coefficients for simulating saola's genetic load, related to Figure 6

(A) Joint distribution of the selection and dominance coefficients used in the (A) Pérez-Pereira model and (B) Kyriazis model. The scatter plot shows the relationship between dominance and selection coefficients, while the histograms in the axes show the corresponding marginal distributions.

(B) Trajectories of population size (top), realized genetic load (middle), and genetic load (bottom) for the northern (left) and southern (right) populations under the inferred saola demographic history and assuming the Pérez-Pereira model of selection and dominance coefficients.

(C) Trajectories of realized genetic load and genetic load for the northern (left) and southern (right) populations under the inferred saola demographic history and assuming the Kyriazis model of selection and dominance coefficients.

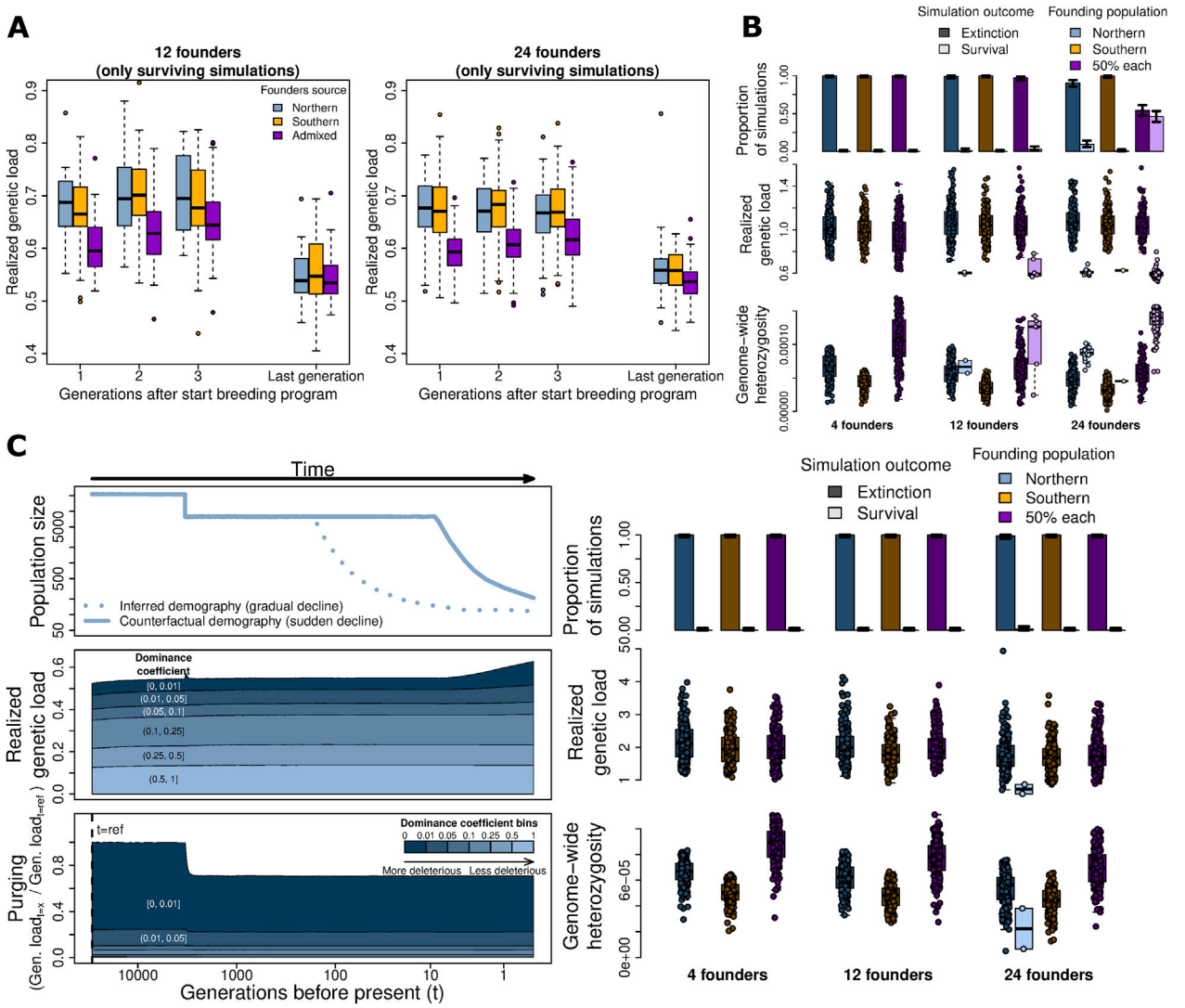


Figure S9. Supplemental results for simulations of genetic load in saola, related to Figure 6

(A) Realized genetic load across the simulation runs that ultimately result in survival in the scenarios for which there were at least 20 simulations resulting in survival. For each combination of number of founders and source population(s) of the founders, the realized genetic load per simulation runs in the three first generations and in the last generations are plotted (note the last simulation is either the 100th generation or the generation when population size grows above 1,000, so it will be variable across runs). In all cases, an admixed breeding program starts with a reduced realized genetic load due to masking of deleterious mutations. Despite the realized load increasing in the succeeding generations, on average it does not go above the realized genetic load in the single population source scenarios, and in the last generation tends to be equal to or lower than. The results are based on simulations using the Pérez-Pereira model of dominance and selection coefficients.

(B) Same as Figure 6B, but using the Kyriazis model for the joint distribution of selection and dominance coefficients.

(C) Same as Figure 6, but showing the results of simulations of genetic load in a counterfactual scenario where saola's population decline had been more recent and sudden, starting 10 generations ago.