Ancient and modern genomes unravel the evolutionary history of the rhinoceros family

Graphical abstract

Blue lineages depict taxa with two horns, whereas red depicts those with only a nasal horn and black stands for the Siberian unicorn (frontal horn)

Authors

Shanlin Liu, Michael V. Westbury, Nicolas Dussex, ..., Pierre-Olivier Antoine, Love Dalén, M. Thomas P. Gilbert

Correspondence

shanlin.liu@cau.edu.cn (S.L.),
love.dalen@nrm.se (L.D.),
tgilbert@sund.ku.dk (M.T.P.G.)

In brief

The comparison of *de novo* genomes from the white, black, Sumatran, and greater one-horned rhinoceroses with the genomes of a historic Javan rhinoceros and three extinct Pleistocene species resolves the evolutionary relationships within the Rhinocerotidae family and reveals that low genetic diversity is a long-term feature of rhinoceroses.

Highlights

- Analysis of genomes from all five extant and three extinct rhinoceros species

- Strong phylogenomic support for the geographical hypothesis of rhinoceros evolution

- Basal split between African and Eurasian lineages in the early Miocene (~16 mya)

- While all rhinoceroses have low genome diversity, it is lowest in modern-day ones

Cell

Article

Ancient and modern genomes unravel the evolutionary history of the rhinoceros family

Shanlin Liu,[1,2,37,*] Michael V. Westbury,[2] Nicolas Dussex,[3,4,5] Kieren J. Mitchell,[6] Mikkel-Holger S. Sinding,[2] Peter D. Heintzman,[7] David A. Duchêne,[2] Joshua D. Kapp,[8] Johanna von Seth,[3,4,5] Holly Heiniger,[6] Fátima Sánchez-Barreiro,[2] Ashot Margaryan,[2] Remi André-Olsen,[9] Binia De Cahsan,[2]

(Author list continued on next page)

[1]Department of Entomology, College of Plant Protection, China Agricultural University, Beijing 100193, China
[2]The GLOBE Institute, Faculty of Health and Medical Sciences, University of Copenhagen, 1353 Copenhagen, Denmark
[3]Centre for Palaeogenetics, Svante Arrhenius vag 20C, Stockholm 10691, Sweden
[4]Department of Bioinformatics and Genetics, Swedish Museum of Natural History, Stockholm 10405, Sweden
[5]Department of Zoology, Stockholm University, Stockholm 10691, Sweden
[6]Australian Centre for Ancient DNA, School of Biological Sciences, University of Adelaide, Adelaide 5005, Australia
[7]The Arctic University Museum of Norway, UiT The Arctic University of Norway, Tromsø 9037, Norway
[8]Department of Ecology and Evolutionary Biology, University of California, Santa Cruz, Santa Cruz, CA 95064, USA
[9]Science for Life Laboratory, Department of Biochemistry and Biophysics, Stockholm University, 17121 Solna, Sweden
[10]China National Genebank, BGI Shenzhen, Shenzhen 518083, China
[11]Center for Ecological and Environmental Sciences, Northwestern Polytechnical University, Xi'an 710072, China
[12]Department of Cell and Molecular Biology, National Bioinformatics Infrastructure Sweden, Science for Life Laboratory, Uppsala University, Uppsala, Sweden
[13]Department of Zoology, University of Venda, Thohoyandou 0950, Republic of South Africa
[14]Editor of the Rhino Resource Center, Utrecht, the Netherlands
[15]School of Biosciences, Sir Martin Evans Building, Cardiff University, Cardiff CF10 3AX, UK
[16]Sustainable Places Research Institute, Cardiff University, Cardiff CF10 3BA, UK
[17]San Diego Zoo Wildlife Alliance, Beckman Center for Conservation Research, San Diego, CA 92027, USA
[18]Rotterdam Zoo, Rotterdam, the Netherlands
[19]N.A. Shilo North-East Interdisciplinary Scientific Research Institute, Far East Branch, Russian Academy of Sciences (NEISRI FEB RAS), Magadan 685000, Russia

(Affiliations continued on next page)

SUMMARY

Only five species of the once-diverse Rhinocerotidae remain, making the reconstruction of their evolutionary history a challenge to biologists since Darwin. We sequenced genomes from five rhinoceros species (three extinct and two living), which we compared to existing data from the remaining three living species and a range of outgroups. We identify an early divergence between extant African and Eurasian lineages, resolving a key debate regarding the phylogeny of extant rhinoceroses. This early Miocene (~16 million years ago [mya]) split post-dates the land bridge formation between the Afro-Arabian and Eurasian landmasses. Our analyses also show that while rhinoceros genomes in general exhibit low levels of genome-wide diversity, heterozygosity is lowest and inbreeding is highest in the modern species. These results suggest that while low genetic diversity is a long-term feature of the family, it has been particularly exacerbated recently, likely reflecting recent anthropogenic-driven population declines.

INTRODUCTION

Understanding the relationships among rhinoceros species and when they diverged has been a question addressed by evolutionary biologists since the dawn of the field. Darwin himself discussed the topic in 1842 as one of a handful of examples in his short treatise on evolution that preceded *On the Origin of Species* in 1859 (Darwin, 1909). Although rhinoceroses were once a diverse clade, extant rhinoceroses comprise only five species, all of which are highly endangered and global priorities for conservation. Rhinocerotoidea, the clade including the rhinoceros family (Rhinocerotidae), diverged from tapirs 55–60 million years ago (mya) in either Eurasia or North America (Bai et al., 2020). The family subsequently radiated into at least 100 species distributed across Africa, Eurasia, North, and Central America (Cerdeño, 1998) and included some of the largest land mammals that ever lived. Most rhinocerotids went extinct prior to the Pleistocene, with just nine species surviving into the Late Pleistocene,

CellPress
OPEN ACCESS

Cell
Article

Guanliang Meng,[10] Chentao Yang,[10] Lei Chen,[11] Tom van der Valk,[12] Yoshan Moodley,[13] Kees Rookmaaker,[14] Michael W. Bruford,[15,16] Oliver Ryder,[17] Cynthia Steiner,[17] Linda G.R. Bruins-van Sonsbeek,[18] Sergey Vartanyan,[19] Chunxue Guo,[10] Alan Cooper,[20] Pavel Kosintsev,[21,22] Irina Kirillova,[23] Adrian M. Lister,[24] Tomas Marques-Bonet,[25,26,27,28] Shyam Gopalakrishnan,[2] Robert R. Dunn,[2,29] Eline D. Lorenzen,[2] Beth Shapiro,[8,30] Guojie Zhang,[10,31,32,33] Pierre-Olivier Antoine,[34] Love Dalén,[3,4,5,36,*] and M. Thomas P. Gilbert[2,35,36,*]

[20]South Australian Museum, Adelaide, SA 5000, Australia
[21]Institute of Plant and Animal Ecology, Ural Branch of the Russian Academy of Sciences, Yekaterinburg, Russia
[22]Ural Federal University, Yekaterinburg, Russia
[23]Institute of Geography, Russian Academy of Sciences, Moscow 119017, Russia
[24]Department of Earth Sciences, Natural History Museum, London, UK
[25]Institute of Evolutionary Biology (UPF-CSIC), Barcelona, Spain
[26]Centre Nacional d'Anàlisi Genòmica, Centre for Genomic Regulation (CNAG-CRG), The Barcelona Institute of Science and Technology, Barcelona, Spain
[27]Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain
[28]Institut Català de Paleontologia Miquel Crusafont, Universitat Autònoma de Barcelona, Barcelona, Spain
[29]Department of Applied Ecology, North Carolina State University, Raleigh, NC, USA
[30]Howard Hughes Medical Institute, University of California, Santa Cruz, Santa Cruz, CA 96050, USA
[31]Villum Center for Biodiversity Genomics, Section for Ecology and Evolution, Department of Biology, University of Copenhagen, Copenhagen, Denmark
[32]State Key Laboratory of Genetic Resources and Evolution, Kunming Institute of Zoology, Chinese Academy of Sciences, Kunming 650223, China
[33]Center for Excellence in Animal Evolution and Genetics, Chinese Academy of Sciences, Kunming 650223, China
[34]Institut des Sciences de l'Évolution, Université Montpellier, CNRS, IRD, EPHE, Montpellier 34095, France
[35]Norwegian University of Science and Technology (NTNU) University Museum, Trondheim 7012, Norway
[36]Senior author
[37]Lead contact
*Correspondence: shanlin.liu@cau.edu.cn (S.L.), love.dalen@nrm.se (L.D.), tgilbert@sund.ku.dk (M.T.P.G.)
https://doi.org/10.1016/j.cell.2021.07.032

during which additional extinctions occurred. These consist of the five extant species, as well as the now-extinct Siberian unicorn (*Elasmotherium sibiricum*), Merck's rhinoceros (*Stephanorhinus kirchbergensis*) and its relative *Stephanorhinus hemitoechus* (not studied here), and the woolly rhinoceros (*Coelodonta antiquitatis*).

Despite decades of study, fundamental questions remain regarding the evolutionary relationships among the extant rhinoceros species and their recently extinct relatives. Furthermore, several rhinoceros species lack available genomic resources that would allow applications, including DNA-based monitoring, conservation management, and environmental DNA studies. To address these questions and needs, we analyzed a genome dataset representing eight rhinoceros species (Figure 1), including all seven genera that survived into the Late Pleistocene (Cerdeño, 1998). Our data include the five extant rhinoceros species represented by four *de novo* genome assemblies of black (*Diceros bicornis*), white (*Ceratotherium simum*), Sumatran (*Dicerorhinus sumatrensis*), and greater one-horned (*Rhinoceros unicornis*, also known as Indian) rhinoceroses and a resequenced genome of a Javan rhinoceros (*R. sondaicus*). The Javan rhinoceros genome was retrieved from a museum specimen dating to 1838 and resequenced to high coverage (25×). In addition, we sequenced the genomes of three extinct rhinoceros species from Late Pleistocene fossils that are close to, or beyond, the radiocarbon dating limit of ~50 thousand years ago (kya), specifically a Siberian unicorn, a Merck's rhinoceros, and a woolly rhinoceros, sequenced to 9×, 12×, and 35× coverage, respectively (Table S1).

RESULTS

Resolving the rhinoceros phylogeny

Three hypotheses have been proposed to explain the phylogenetic relationships within living Rhinocerotidae: (1) the "horn hypothesis," which groups the two-horned rhinoceros species together, specifically placing the Sumatran rhinoceros as sister to the African Diceroti (black and white rhinoceroses) and has been supported by morphology (Antoine et al., 2010), genetic (e.g. Steiner and Ryder, 2011), and paleoproteomic analyses of dental enamel (Cappellini et al., 2019); (2) the "geographical hypothesis," which places the Asian species together, with Sumatran as sister to the greater one-horned and Javan rhinoceroses, and is based on morpho-anatomical evidence (Antoine et al., 2021), biogeographic parsimony, genetic analyses using a limited number of loci (Kirillova et al., 2017; Kosintsev et al., 2019; Orlando et al., 2003; Tougard et al., 2001), and paleoproteomic analysis using collagen sequences (Welker et al., 2017); and (3) a hypothesis that the Sumatran rhinoceros is sister to the clade comprising the four other extant species, which has been supported by a more recent analysis of complete mitochondrial genomes (Margaryan et al., 2020). These conflicting hypotheses emphasize the limitations of using lower-resolution markers in reconstructing evolutionary relationships within Rhinocerotidae and highlight the potential of applying phylogenomic approaches. Prior studies have also debated the phylogenetic placement of the three extinct species included in this study. For example, the relationship of Merck's and woolly rhinoceroses to each other, the Sumatran rhinoceros, and the two African Diceroti remained contentious due to contrasting
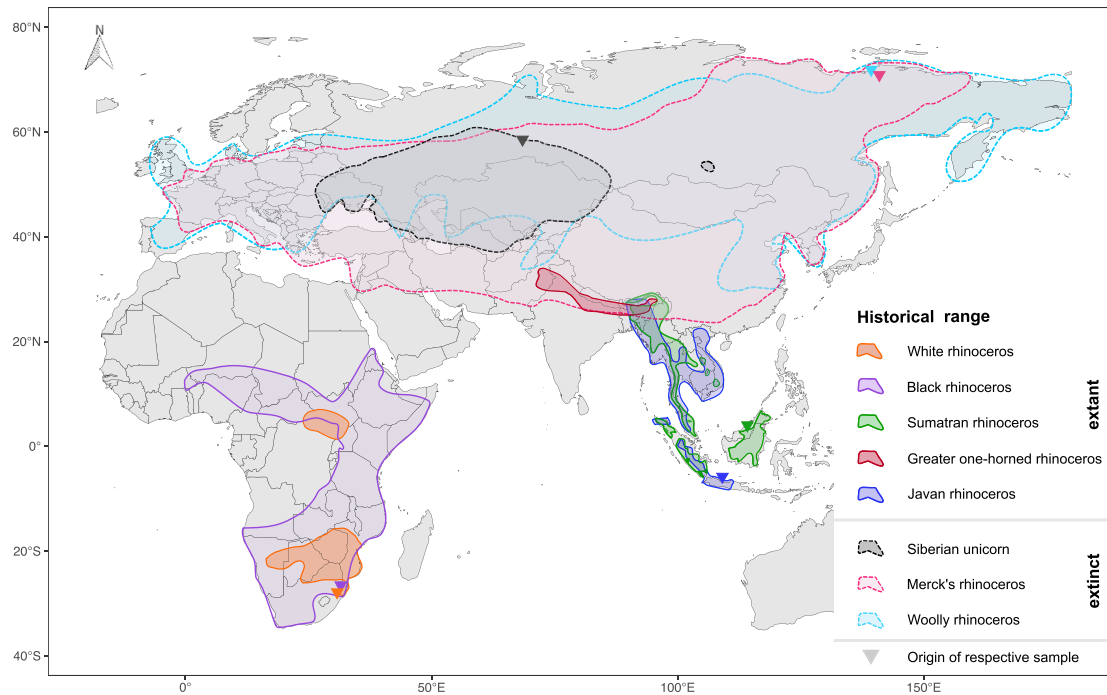
Figure 1. Ranges of the eight rhinoceros species studied

The historical distribution range maps of the five extant species use information published previously (Dinerstein, 2003; Havmøller et al., 2016; Khan and van Strien, 1997; Rookmaaker and Antoine, 2012), while the ranges of the three extinct species were drawn based on their fossil records (Kahlke and Lacombat, 2008; Kosintsev et al., 2019, 2020; Shpansky and Boeskorov, 2018). All ranges are approximations aimed at conveying each species' general, rather than detailed, distribution. As the greater one-horned rhinoceros sample derives from captive-born zoo stock, a geographic origin is not shown.

conclusions drawn from morpho-anatomical versus mitogenomic and paleoproteomic evidence (Antoine et al., 2021; Cappellini et al., 2019; Kirillova et al., 2017; Kosintsev et al., 2019).

To circumvent the problem of reference genome biases (Gopalakrishnan et al., 2017; Heintzman et al., 2017; Figures S1A and S1B), we conducted whole-genome alignments for the eight rhinoceros species using domestic horse (*Equus caballus*) as an outgroup. We then inferred a genome-wide species tree summarizing the phylogenetic signal from individual gene trees based on 22,066 100-kb genomic windows. Our phylogenomic analysis identified three major clades within the subfamily Rhinocerotinae and provided strong support for the geographical hypothesis of rhinoceros evolution. A clade comprising the two African species *Diceros bicornis* and *Ceratotherium simum*, the Diceroti, is the sister lineage to the remaining rhinoceroses in our dataset (with the exception of the Siberian unicorn). A second clade includes the Sumatran, Merck's, and woolly rhinoceroses (hereafter referred to as the *Dicerorhinus-Coelodonta-Stephanorhinus* [*DCS*] clade), all of which have current or past geographic ranges that include parts of Asia. The third clade includes the two *Rhinoceros* species (Figure 2A). Thus, the principal divergence among the rhinoceros lineages is related to the geographical division between species on the African and Eurasian continents.

Our phylogenomic analyses confirm prior conclusions based on morphological and biomolecular evidence (Figure 2A) that place the extinct Siberian unicorn as outgroup to the subfamily Rhinocerotinae (P.-O.A., unpublished data; Becker et al., 2013;

Kosintsev et al., 2019). Within the *DCS* clade, we found strong support for the Sumatran rhinoceros as sister to the clade that includes the extinct Merck's and woolly rhinoceroses. A remaining challenge is understanding the relationships of more ancient extinct species, for which DNA remains unrecoverable. As our findings suggest that results based purely on morphology (the horn hypothesis) are not supported, attempting to fill in the rest of the phylogeny based solely on the morphology of extinct taxa may prove difficult.

Gene flow among species

While we were able to resolve a genome-wide rhinoceros species tree, we uncovered significant phylogenetic discordance across the rhinoceros genomes, suggesting gene flow or incomplete lineage sorting (ILS) among taxa. While this topology represents the dominant signal of species relationships across the whole genome, analyses of individual chromosomes did not always recover the same topology (Figure S1C). Most prominently, we observed that the species-tree position of the *DCS* clade was supported by only ~45% of the individual sliding-window trees, which is substantially fewer than recovered for other nodes in the phylogeny (Figure 2B). Because a genomic region of 100 kb in length may contain multiple recombination breakpoints, we also inferred gene trees using 5-kb alignments randomly subsampled from within each 100-kb sliding window. The results corroborated the phylogenetic discordance discovered in the initial dataset (Figure S1D). We then simulated the gene tree
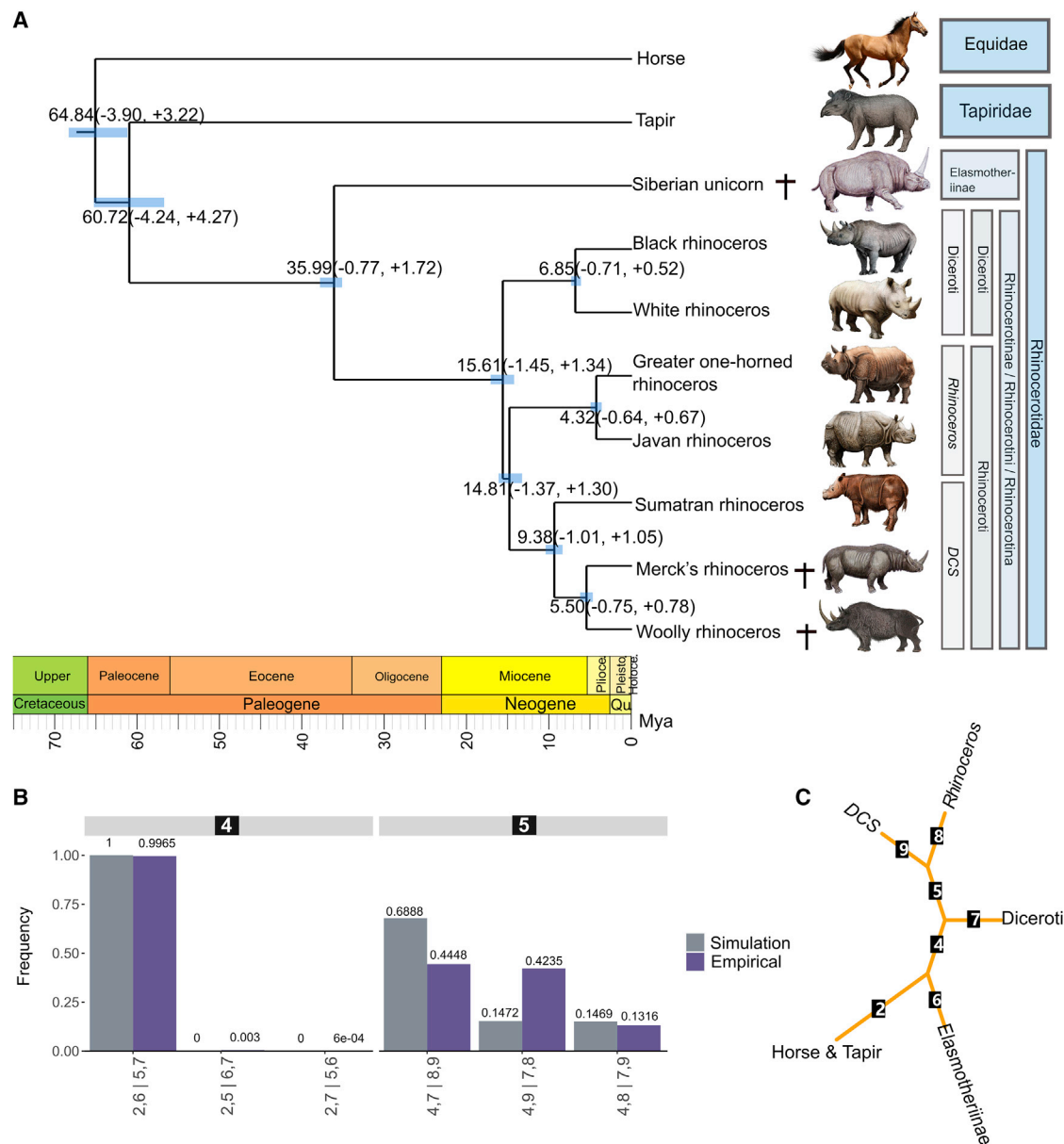
Figure 2. Phylogeny of the Rhinocerotidae
(A) Dated species tree of the Rhinocerotidae based on a consensus of trees generated every 100 kb across the genome using maximum likelihood methods and multiple fossil calibrations as detailed in STAR Methods. Blue horizontal bars show 95% confidence intervals of estimated divergence dates between lineages. Black crosses indicate extinct species.
(B and C) Frequency of the three bipartitions induced by branches 4 and 5 (B) in the tree (C). Numbers on the x axis text and header (B) correspond to the branch identity of the tree (C). For example 4,7|8,9 represents the bipartition ((horse and tapir, *Elasmotherium*) Diceroti) | (*Rhinoceros*, DCS) induced by branch 5.

distribution expected under a hypothesis of genetic drift with ILS and compared this to the empirical data. We found disagreement between the simulated and empirical data, suggestive of both ILS and gene flow as the cause of the diminished support for the species tree. We found in the empirical data an excess of gene tree topologies supporting the placement of the Sumatran rhinoceros as sister to the other species of the Rhinocerotinae subfamily, as observed previously in some mitochondrial phy-

logenies (Willerslev et al., 2009; Margaryan et al., 2020; Figure 2B).

To further explore the origins of the phylogenomic discordances, we used D-statistics analyses to differentiate ancient gene flow events from ILS (Green et al., 2010). The use of this method can be problematic when applied to relatively divergent species due both to possible violations in assumptions such as equal mutation rates and the infinite sites model and biases

Cell
Article

Figure 3. Comparison of whole-genome heterozygosity estimated in various taxa and ROH distribution in the eight rhinoceros species

(A) Heterozygosity estimates of a broad range of animals, mammals, and ruminants. Species with heterozygosity values >1% were not included. See also Table S3.

(B) Runs of homozygosity (ROH) size distributions for seven of the species investigated. We did not identify any ROH in the Siberian unicorn; data are not shown here, as we cannot exclude the influence of reference bias.

derived from the mapping of resequenced data to a reference genome (Figures S2A and S2B). Despite these potential sources of error, the D-statistics are congruent with the phylogenomic analyses in suggesting that gene flow and ILS occurred between the ancestors of the Diceroti and *Rhinoceros* species (Figures 2B and 2C). This gene flow may have been enabled by the Eurasian origin of both African species (Antoine et al., 2021; Geraads, 2005, 2020). Our analyses also revealed an excess of shared derived alleles between the two extinct members of the *DCS* clade (Merck's and woolly rhinoceroses) and both representatives of *Rhinoceros*, suggesting gene flow between these two pairs of lineages. We found no evidence of excess shared derived alleles between the Sumatran rhinoceros and either *Rhinoceros* species (Figure S2D), despite their closer geographic proximity (Figure 1). This suggests either no gene flow or similar levels of gene flow between the Sumatran rhino and both *Rhinoceros* species. These differential patterns of gene flow may explain the discrepancy of the phylogenetic placement of the Sumatran rhinoceros as sister to the genus *Rhinoceros* based on nuclear DNA versus sister to all other members of the subfamily Rhinocerotinae in some mitochondrial phylogenies (Willerslev et al., 2009; Margaryan et al., 2020).

Timing of divergence between species

We used fossil data to calibrate our phylogeny and estimate lineage divergence times (Figure 2A). This resulted in a ∼65 mya estimate for the common ancestor of horses, tapirs, and rhinoceroses and a ∼36 mya estimate for the common ancestor of the extinct rhinoceros subfamily Elasmotheriinae and the extant subfamily Rhinocerotinae. The three major clades within the Rhinocerotinae subfamily diverged ∼16 mya (Figure 2A), at the end of the early Miocene and around the time of the Miocene climatic optimum (17–14 mya), a period that was ∼3°C–4°C warmer than present (Lewis et al., 2008; Sosdian et al., 2020). Diversification occurred after the formation of the terrestrial connection between the Afro-Arabian and Eurasian landmasses ∼20 mya (Van Couvering and Delson, 2020). We hypothesize that this land bridge enabled dispersal events followed by vicariance, as is well documented with the immigration into Africa from Eurasia of early rhinocerotids, giraffids, suids, and viverrids and the emigration from Africa to Eurasia of apes, deinotheres, and elephantoids, among others (Van Couvering and Delson, 2020).

Drivers of low genetic diversity

Previous genomic studies on black, white, and Sumatran rhinoceroses identified low levels of genetic diversity (Mays et al., 2018; Moodley et al., 2020; Tunstall et al., 2018). These findings are consistent with the observation that all extant rhinoceros species have gone through recent population size declines, even though some species (white and greater one-horned rhinoceroses) have since recovered (Ellis and Talukdar, 2019; Emslie, 2020; Rookmaaker and Antoine, 2012). However, low genetic diversity can also be a consequence of particular life-history traits and/or long-term small population size (Westbury et al., 2018, 2019; Xue et al., 2015). To investigate this, we calculated genome-wide heterozygosity (GWH) for all eight rhinoceros species and compared these estimates with GWH in a range of other animal species, including ruminants and, more broadly, mammals. We assessed whether GWH levels are lower in genomes recovered from present-day animals (i.e., black, white, greater one-horned, and Sumatran rhinoceroses) compared to GWH in genomes recovered from specimens that pre-date the human-mediated declines during the last 100 years (i.e., the nearly 180-year-old Javan rhinoceros genome as well as the genomes from the three extinct species).

We estimated GWH based on transversions only to limit the potential influence of DNA damage on estimates from the ancient and historical genomes. However, for comparability with published results for other taxa, which incorporate all variable sites, we recalibrated our estimates based on the expected transition/transversion ratio (see Figure S3A). Our results showed that present-day rhinoceros genomes exhibit significantly lower GWH compared to the historical Javan and extinct genomes (one-way ANOVA, n = 8, F = 7.4, p = 0.04). On the other hand, our comparison with a broad range of animals shows that rhinoceroses in general display comparatively low levels of GWH, especially relative to not only the combined dataset of all animals but also ruminants and other large herbivores (Figure 3A). The only mammalian family displaying lower average levels of GWH was the Felidae (Figure S3B), which is not unexpected, as carnivores/predators are generally less abundant than herbivores/prey (Owen-Smith, 2015). These findings are robust to choice of reference genome used in our analyses (see Figure S3C).

To better contextualize the observed levels of GWH, we characterized the inbreeding levels in our genomes through

distributions of runs of homozygosity (ROH). To evaluate the robustness of our results, we first explored the effect of excluding transitions on ROH inference (Figure S4A). This analysis indicated that a reliance solely on transversions shifts the distribution of ROH segments to longer stretches but that this is only a problem for species with low overall GWH. Since all the resequenced genomes (Javan, Siberian unicorn, Merck's, and woolly rhinoceroses) exhibit higher GWH than at least three of the *de novo* assembled genomes (black, white, Sumatran, and greater one-horned rhinoceros), a reliance solely on transversions for these individuals may somewhat artificially inflate the length of ROH segments but should not bias our overall interpretations.

We did not detect any ROH segments in the Siberian unicorn, which may reflect its exceptional phylogenetic distance to the white rhinoceros genome against which it was mapped, artificially inflating heterozygous sites (Figure S3C). In contrast, ROH segments >2 Mb were detected in all other rhinoceros species, where all species except the woolly rhinoceros also exhibited stretches as long as 5 Mb. However, we observed significantly higher inbreeding levels (one-way ANOVA, n = 7, F = 36.7, p = 0.002), which are proportional to the overall length of ROH segments, in the genomes from present-day rhinoceroses compared to the genomes of the historical Javan and extinct Merck's and woolly rhinoceroses (Figure 3B).

Overall, the comparisons of GWH and inbreeding levels suggest that recent population declines caused by heavy anthropogenic pressure in the 20th century (e.g., Player, 1973) resulted in marked losses in genetic diversity, as well as increased inbreeding levels. However, the genomes from historical and extinct species, which were sampled either prior to their recent population collapse or many thousands of years before their extinction, also exhibit low levels of GWH when compared to other animal species (Figure 3A). Moreover, the observation of a moderate amount of long ROH segments in the genomes from the Javan, Merck's, and woolly rhinoceroses is consistent with background inbreeding in these species. We thus hypothesize that limited genetic diversity and moderate inbreeding levels are intrinsic features of rhinoceros life history, where low population densities and limited dispersal result in increased genetic drift as well as occasional mating between relatives.

Demography and mutational load

To further assess the genomic background of the overall low GWH and moderate inbreeding levels in Rhinocerotidae, we modeled changes in effective population size (N_e) throughout the Pleistocene using the pairwise sequentially Markovian coalescent (PSMC) model (Li and Durbin, 2011; Figure 4). Although previous studies have reconstructed the demographic histories for a subset of species, including black rhinoceros (Moodley et al., 2020), white rhinoceros (Tunstall et al., 2018), Sumatran rhinoceros (Mays et al., 2018), and woolly rhinoceros (Lord et al., 2020), our combined analysis allows exploration of both shared and unique responses through time. Overall, it is striking that all eight species displayed either a general continual decrease in N_e over the last two million years or a continuously small N_e over extended time periods.

Previous studies have suggested that maintenance of a low population size over extended periods of time allows for the purging of deleterious alleles while keeping low genome-wide levels of genetic diversity (Westbury et al., 2018, 2019; Xue et al., 2015). Our finding that all rhinoceros species have had a small N_e during extended periods of their history could thus indicate a similar scenario.

To investigate how mutational load in Rhinocerotidae compares to other species, we calculated ratios of loss-of-function (LoF; generally highly deleterious) versus synonymous mutations for seven of the rhinoceros species and 30 mammalian species from diverse clades (Table S4). The results show that the levels of mutational load in rhinoceroses fall within the range observed among other present-day mammals (Figure 5). However, the extinct and historic rhinoceros genomes (Siberian unicorn, woolly, and Javan rhinoceroses) displayed a significantly higher number of LoF mutations (Figure 5; one-way ANOVA, n = 7, F = 29.0, p = 0.003) compared to the present-day rhinoceros genomes (i.e., black, white, Sumatran, and greater one-horned rhinoceroses). We thus find no evidence for an accumulation of mutational load within the last decades for those species that have gone through recent population declines. Although speculative, we therefore hypothesize that extant rhinoceroses may have undergone some purging of mutational load in connection with their demographic declines in the last 100 years. However, this hypothesis requires further testing, for example through intra-specific comparisons of historical and modern genomes that span these declines and, in some cases, recoveries (Sánchez-Barreiro et al., 2021; von Seth et al., 2021). Meanwhile, it is worth noting that some of the species used in the analysis were mapped onto their closely related species with available reference genomes, and the qualities of genome assemblies and annotations vary between species as well, both of which could influence the accuracy of gene effect estimation.

DISCUSSION

Our combined rhinoceros genome dataset represents a valuable resource for both the current and future study of the evolution and biology of these species, including characterization of the genetic basis of rhinoceros phenotypes (Table S5). For example, we uncovered frameshift mutations in IFT43 (intraflagellar transport 43) that could contribute to rhinoceroses' poor eyesight. IFT43 is involved in the formation and maintenance of cilia, which are important for the development and function of the light-sensitive tissue at the back of the eye (the retina) (Arts et al., 2011).

In the case of the Javan and greater one-horned rhinoceros, our genome sequences also provide a basis for further species-specific conservation genetic analyses. By analyzing this dataset, we resolved a long-standing debate related to the evolutionary history of living and recently extinct rhinoceroses and provided evidence that relatively low genomic heterozygosity and moderate inbreeding levels may represent their long-term natural state. These findings suggest that low levels of diversity and high inbreeding observed in present-day rhinoceros genomes can only partially be attributed to recent declines. This may be positive news for conservation, since it implies that recent declines may have had less impact on the genetic aspects of population
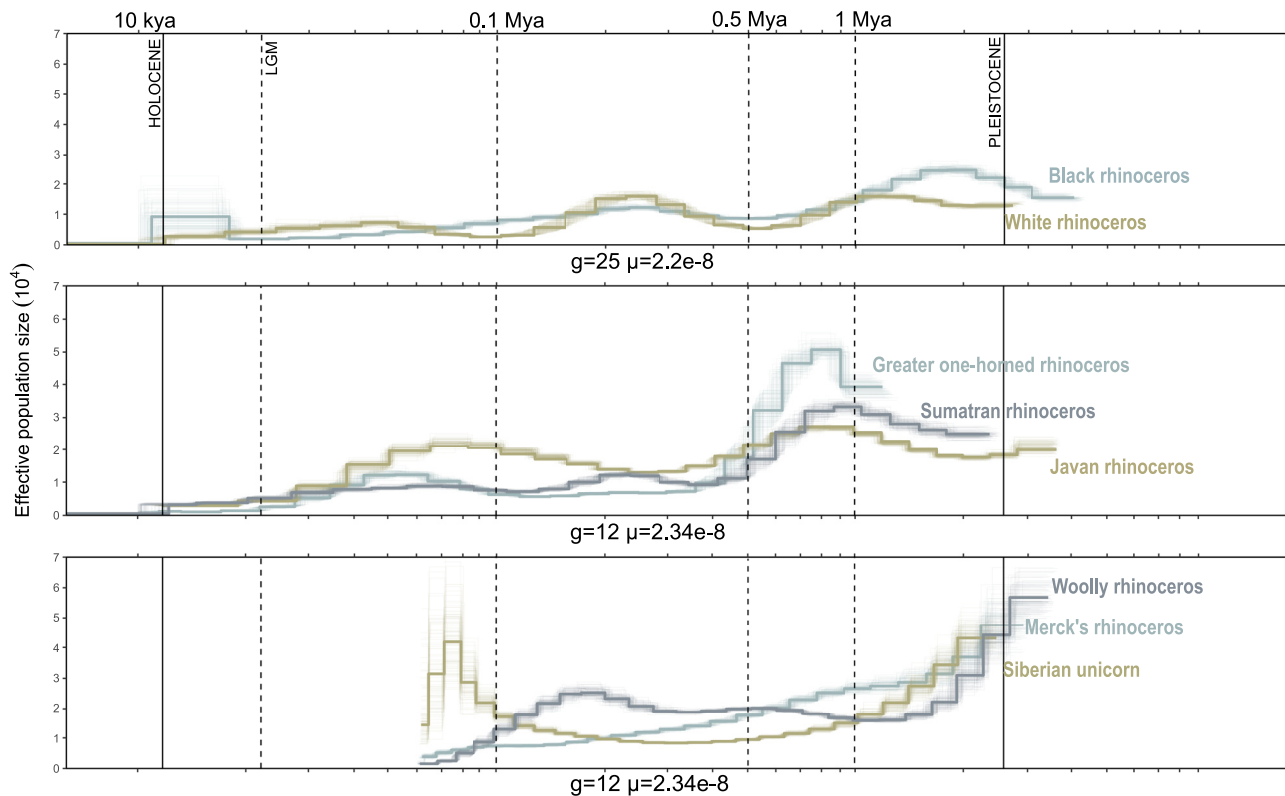
Figure 4. Demographic trajectory of the eight rhinoceros species
Each curve represents one species, with thin lines depicting bootstraps. The x axis corresponds to time before present in years on a log scale. LGM, last glacial maximum. We assumed two different substitution rates (substitutions/site/generation, μ) and generation times (g) for our rhinoceros species and set the sample age of the three extinct species to 50,000 years (see Table S1). The y axis corresponds to the effective population size (N_e). Species are grouped by geographic distribution: Africa, black and white rhinoceroses; South Asia, Sumatran, Javan, and greater one-horned rhinoceroses; and northern Eurasia, Siberian unicorn, Merck's, and woolly rhinoceroses.

viability than previously thought. Nonetheless, extant rhinoceroses undoubtedly face enormous challenges in the future, principally due to anthropogenic and environmental effects. A major priority for rhinoceros conservation will be to halt illegal poaching and ensure that there is sufficient carrying capacity for population recovery. Our study highlights how genomics can complement such actions by enabling monitoring of ongoing changes in genetic variation, inbreeding, and mutational load.

Limitations of the study
Given the historic and ancient nature of the specimens from four of the species studied, their DNA quality was not suitable for *de novo* assembly; thus, their genome sequences were recovered through mapping against other species. This process can introduce biases in downstream analyses that can arise due to differential mapping efficiencies influenced by phylogenetic distance to the reference genome, ancient DNA damage, and short read lengths. Therefore, although we took several steps to alleviate such influences, we highlight that this should be kept in mind. Furthermore, it should be noted that the rhinoceros family was once a speciose group, and only a small fraction of these were studied here. Thus, given that we lack genome sequences from the majority of the clade, clearly a huge gap will remain

that needs to be bridged before we can fully understand the evolutionary history of the rhinoceros family.

STAR★METHODS

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

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
  - Source organisms
- METHOD DETAILS
  - DNA isolation and sequencing
  - Genome assembly and annotation
  - Genome alignment and phylogenetic inference
  - Molecular dating
  - Alignment tool selection
  - Gene flow
  - Heterozygosity estimation
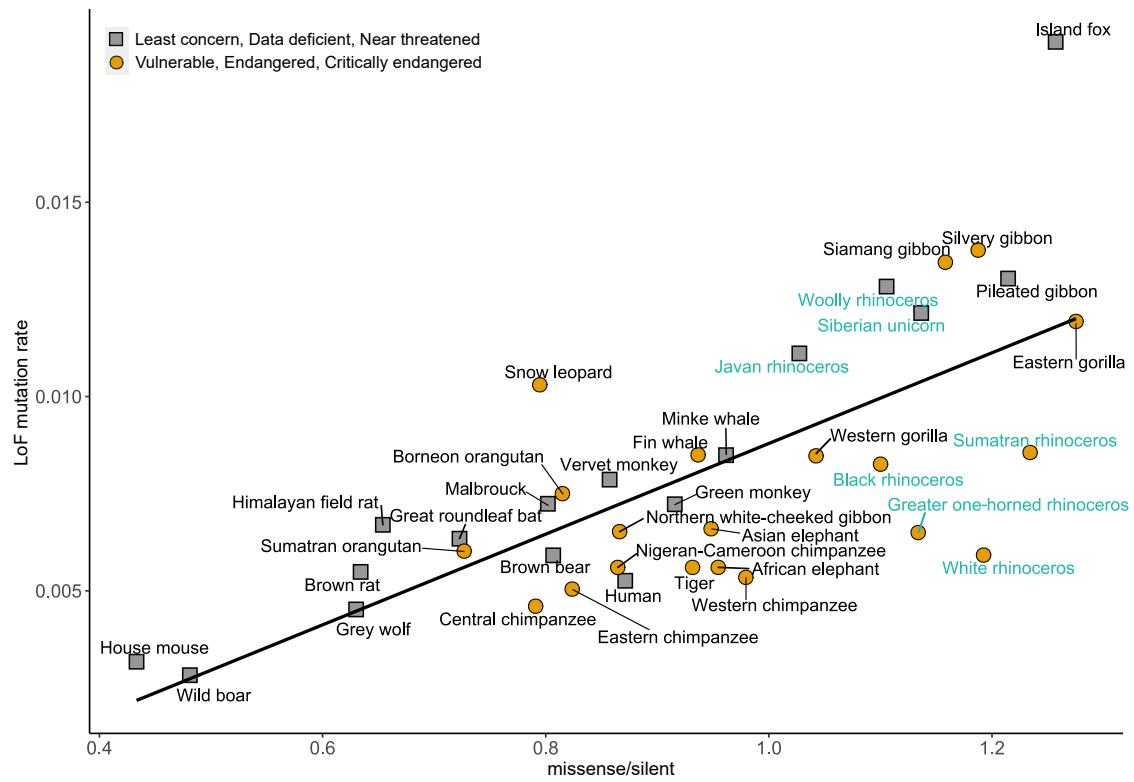
Figure 5. Genome variant effect of a broad range of mammal species
The black line represents a linear regression fit that showed a significant correlation between loss-of-function (LoF) mutation rate and rate of missense/silent ($r = 0.72$, $p = 4.05e-7$). Note that the Javan rhinoceros sample dates to 1838, so we set its conservation status as data deficient. The rhinoceros species are shown in turquoise lettering. LoF mutations are here represented by nonsense mutations, and the rate is used as a proxy for the accumulation of mutation load. Missense/silent means the rate of number of missense mutations relative to that of silent mutations. For species with multiple individuals, this figure shows their mean values. Note that the result of Merck's rhinoceros was not included due to its abnormal transversion/transition ratio (Figure S4B).

○ Demographic inference and Runs of Homozygosity (ROH) estimation
○ Genetic load and identification for rhinoceros specific frameshift mutation
○ Test of heterozygosity, PSMC and ROH estimation for the non-modern rhino samples
● QUANTIFICATION AND STATISTICAL ANALYSIS
● ADDITIONAL RESOURCES

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.cell.2021.07.032.

ACKNOWLEDGMENTS

AUTHOR CONTRIBUTIONS

S.L., L.D., and M.T.P.G. conceived the project and designed the research. M.-H.S.S., K.J.M., S.V., P.K., I.K., A.C., B.S., and G.Z. provided archaeological work, logistics, and/or ancient samples and data. F.S.-B., Y.M., M.B., T.v.d.V., O.R., C.S., and L.G.R.B.-v.S. coordinated logistics of and/or provided modern samples and data. M.-H.S.S., N.D., K.J.M., P.D.H., J.D.K., J.v.S., H.H., C.G., G.M., and C.Y. conducted laboratory work. S.L., R.A.-O., D.A.D., M.V.W., and L.C. conducted analyses of data with considerable input from A.M., T.v.d.V., S.G., P.D.H., T.M.-B., P.-O.A., L.D., and M.T.P.G. S.L., M.V.W., N.D., K.J.M., P.D.H., D.A.D., F.S.-B., A.M., B.D.C., Y.M., K.R., A.L., T.M.-B., S.G.,

Cell
Article

CellPress
OPEN ACCESS

SUPPORTING CITATIONS

The following references appear in the supplemental information: Carbone et al. (2014); Chen et al. (2019); de Manuel et al. (2016); Dong et al. (2017); Durand et al. (2011); Freedman et al. (2014); Groenen et al. (2012); Harr et al. (2016); Kardos et al. (2018); Karsten et al. (2011); Liu et al. (2014); Lynch et al. (2015); Mallick et al. (2016); Nater et al. (2017); Palkopoulou et al. (2018); Prado-Martinez (2013); (Reddy et al., 2015); Robinson et al. (2016); Smith et al. (2017); Svardal et al. (2017); Teng et al. (2017); Veeramah et al. (2015); Wang et al. (2019); Wilkie et al. (2013); Yim et al. (2014); Yoo et al. (2017); (Zhang et al., 2014); Zoonomia Consortium (2020); (Cho et al., 2013)

REFERENCES

Antoine, P.-O., Downing, K.F., Crochet, J., Duranthon, F., Flynn, L.J., Marivaux, L., Metais, G., Rajpar, A.R., and Roohi, G. (2010). A revision of Aceratherium blanfordi Lydekker, 1884 (Mammalia: Rhinocerotidae) from the Early Miocene of Pakistan: postcranials as a key. Zool. J. Linn. Soc. 160, 139–194.

Antoine, P.-O., Métais, G., Orliac, M.J., Crochet, J.-Y., Flynn, L.J., Marivaux, L., Rajpar, A.R., Roohi, G., and Welcomme, J.-L. (2013). Mammalian Neogene biostratigraphy of the Sulaiman Province, Pakistan. In Fossil mammals of Asia: Neogene Biostratigraphy and Chronology, X.-m. Wang, L.J. Flynn, and M. Fortelius, eds. (Columbia University Press), pp. 400–422.

Antoine, P.-O., Reyes, M.C., Amano, N., Claude, J., Bautista, A.P., Vos, J.d., and Ingicco, T. (2021). A new clade of rhinoceroses from the Pleistocene of Asia sheds light on mainland mammal dispersals to the Philippines. Zool. J. Linn. Soc. Published online May 22, 2021. https://doi.org/10.1093/zoolinnean/zlab009.

Arts, H.H., Bongers, E.M.H.F., Mans, D.A., van Beersum, S.E., Oud, M.M., Bolat, E., Spruijt, L., Cornelissen, E.A.M., Schuurs-Hoeijmakers, J.H.M., de Leeuw, N., et al. (2011). C14ORF179 encoding IFT43 is mutated in Sensenbrenner syndrome. J. Med. Genet. 48, 390–395.

Auwera, G.A.V.d., Carneiro, M.O., Hartl, C., Poplin, R., Angel, G.d., Levy-Moonshine, A., Jordan, T., Shakir, K., Roazen, D., Thibault, J., et al. (2013). From FastQ data to high-confidence variant calls: the Genome Analysis Toolkit best practices pipeline. Curr. Protoc. Bioinformatics. 43, 11.10.11–11.10.33.

Bai, B., Meng, J., Zhang, C., Gong, Y.-X., and Wang, Y.-Q. (2020). The origin of Rhinocerotoidea and phylogeny of Ceratomorpha (Mammalia, Perissodactyla). Commun. Biol. 3, 509.

Becker, D., Antoine, P., and Maridet, O. (2013). A new genus of Rhinocerotidae (Mammalia, Perissodactyla) from the Oligocene of Europe. J. Syst. Palaeontology 11, 947–972.

Blanchette, M., Kent, W.J., Riemer, C., Elnitski, L., Smit, A.F., 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.

Böhme, M., Aiglstorfer, M., Antoine, P., Appel, E., Havlik, P., Metais, G., Schneider, S., Setzer, F., Tappert, R., and Tran, D.N. (2013). Na Duong (northern Vietnam) – an exceptional window into Eocene ecosystems from Southeast Asia. Zitteliana 53, 121–167.

Brace, S., Palkopoulou, E., Dalén, L., Lister, A.M., Miller, R., Otte, M., Germonpré, M., Blockley, S.P.E., Stewart, J.R., and Barnes, I. (2012). Serial population extinctions in a small mammal indicate Late Pleistocene ecosystem instability. Proc. Natl. Acad. Sci. USA 109, 20532–20536.

Briggs, A.W., Stenzel, U., Johnson, P.L.F., Green, R.E., Kelso, J., Prüfer, K., Meyer, M., Krause, J., Ronan, M.T., Lachmann, M., and Pääbo, S. (2007). Patterns of damage in genomic DNA sequences from a Neandertal. Proc. Natl. Acad. Sci. USA 104, 14616–14621.

Briggs, A.W., Stenzel, U., Meyer, M., Krause, J., Kircher, M., and Pääbo, S. (2010). Removal of deaminated cytosines and detection of in vivo methylation in ancient DNA. Nucleic Acids Res. 38, e87.

Brotherton, P., Endicott, P., Sánchez, J.J.M., Beaumont, M., Barnett, R., Austin, J., and Cooper, A. (2007). Novel high-resolution characterization of ancient DNA reveals C > U-type base modification events as the sole cause of post mortem miscoding lesions. Nucleic Acids Res. 35, 5717–5728.

Brown, F.H., and McDougall, I. (2011). Geochronology of the Turkana depression of northern Kenya and southern Ethiopia. Evol. Anthropol. 20, 217–227.

Buckley, M. (2015). Ancient collagen reveals evolutionary history of the endemic South American 'ungulates'. Proc. Biol. Sci. 282, 20142671, 20142671.

Cappellini, E., Jensen, L.J., Szklarczyk, D., Ginolhac, A., da Fonseca, R.A., Stafford, T.W., Holen, S.R., Collins, M.J., Orlando, L., Willerslev, E., et al. (2012). Proteomic analysis of a pleistocene mammoth femur reveals more than one hundred ancient bone proteins. J. Proteome Res. 11, 917–926.

Cappellini, E., Welker, F., Pandolfi, L., Ramos-Madrigal, J., Samodova, D., Rüther, P.L., Fotakis, A.K., Lyon, D., Moreno-Mayar, J.V., Bukhsianidze, M., et al. (2019). Early Pleistocene enamel proteome from Dmanisi resolves Stephanorhinus phylogeny. Nature 574, 103–107.

Carbone, L., Harris, R.A., Gnerre, S., Veeramah, K.R., Lorente-Galdos, B., Huddleston, J., Meyer, T.J., Herrero, J., Roos, C., Aken, B., et al. (2014). Gibbon genome and the fast karyotype evolution of small apes. Nature 513, 195–201.

Caroe, C., Gopalakrishnan, S., Vinner, L., Mak, S.S.T., Sinding, M.S., Samaniego, J.A., Wales, N., Sicheritzponten, T., and Gilbert, M.T.P. (2017). Single-tube library preparation for degraded DNA. Methods Ecol. Evol. 9, 410–419.

Cerdeño, E. (1998). Diversity and evolutionary trends of the Family Rhinocerotidae (Perissodactyla). Palaeogeogr. Palaeoclimatol. Palaeoecol. 141, 13–34.

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, 364.

Cho, Y.S., Hu, L., Hou, H., Lee, H., Xu, J., Kwon, S., Oh, S., Kim, H.-M., Jho, S., Kim, S., et al. (2013). The tiger genome and comparative analysis with lion and snow leopard genomes. Nat. Commun. 4, 2433.

Cingolani, P., Platts, A., Wang, 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.

Dabney, J., Knapp, M., Glocke, I., Gansauge, M.T., Weihmann, A., Nickel, B., Valdiosera, C., García, N., Pääbo, S., Arsuaga, J.L., and Meyer, M. (2013). Complete mitochondrial genome sequence of a Middle Pleistocene cave bear reconstructed from ultrashort DNA fragments. Proc. Natl. Acad. Sci. USA 110, 15758–15763.

F. Darwin, ed. (1909). The foundations of the origin of species. Two essays written in 1842 and 1844 (Cambridge University Press).

de Manuel, M., Kuhlwilm, M., Frandsen, P., Sousa, V.C., Desai, T., Prado-Martinez, J., Hernandez-Rodriguez, J., Dupanloup, I., Lao, O., Hallast, P., et al. (2016). Chimpanzee genomic diversity reveals ancient admixture with bonobos. Science 354, 477–481.

Dinerstein, E. (2003). The return of the unicorns: the natural history and conservation of the greater one-horned rhinoceros (Columbia University Press).

Dinerstein, E., and McCracken, G.F. (1990). Endangered greater one-horned rhinoceros carry high levels of genetic variation. Conserv. Biol. 4, 417–422.

Dong, D., Lei, M., Hua, P., Pan, Y.-H., Mu, S., Zheng, G., Pang, E., Lin, K., and Zhang, S. (2017). The genomes of two bat species with long constant frequency echolocation calls. Mol. Biol. Evol. *34*, 20–34.

dos Reis, M., Donoghue, P.C.J., and Yang, Z. (2014). Neither phylogenomic nor palaeontological data support a Palaeogene origin of placental mammals. Biol. Lett. *10*, 20131003.

Doyle, V.P., Young, R.E., Naylor, G.J.P., and Brown, J.M. (2015). Can we identify genes with Increased Phylogenetic Reliability. Syst. Biol. *64*, 824–837.

Durand, E.Y., Patterson, N., Reich, D., and Slatkin, M. (2011). Testing for ancient admixture between closely related populations. Mol. Biol. Evol. *28*, 2239–2252.

Ellis, S., and Talukdar, B. (2019). *Rhinoceros unicornis*. In the IUCN red list of threatened species 2019: eT19496A18494149 (IUCN). https://dx.doi.org/10.2305/IUCN.UK.2019-3.RLTS.T19496A18494149.en.

Emslie, R. (2020). *Ceratotherium simum*. In the IUCN red list of threatened species 2020: eT4185A45813880 (IUCN). https://dx.doi.org/10.2305/IUCN.UK.2020-1.RLTS.T4185A45813880.en.

Ersmark, E., Orlando, L., Sandoval-Castellanos, E., Barnes, I., Barnett, R., Stuart, A.J., Lister, A.M., and Dalen, L. (2015). Population demography and genetic diversity in the Pleistocene cave lion. Open Quat. *1*, 1–14.

Freedman, A.H., Gronau, I., Schweizer, R.M., Ortega-Del Vecchyo, D., Han, E., Silva, P.M., Galaverni, M., Fan, Z., Marx, P., Lorente-Galdos, B., et al. (2014). Genome sequencing highlights the dynamic early history of dogs. PLoS Genet. *10*, e1004016.

Geraads, D. (2005). Pliocene Rhinocerotidae (Mammalia) from Hadar and Dikika (Lower Awash, Ethiopia), and a revision of the origin of modern african rhinos. J. Vertebr. Paleontol. *25*, 451–461.

Geraads, D. (2020). Perissodactyla (Rhinocerotidae and Equidae) from Kanapoi. J. Hum. Evol. *140*, 102373.

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.

Gnerre, S., Maccallum, I., Przybylski, D., Ribeiro, F.J., Burton, J.N., Walker, B.J., Sharpe, T., Hall, G., Shea, T.P., Sykes, S., et al. (2011). High-quality draft assemblies of mammalian genomes from massively parallel sequence data. Proc. Natl. Acad. Sci. USA *108*, 1513–1518.

Gopalakrishnan, S., Samaniego Castruita, J.A., Sinding, M.S., Kuderna, L.F.K., Räikkönen, J., Petersen, B., Sicheritz-Ponten, T., Larson, G., Orlando, L., Marques-Bonet, T., et al. (2017). The wolf reference genome sequence (*Canis lupus lupus*) and its implications for Canis spp. population genomics. BMC Genomics *18*, 495.

Green, R.E., Krause, J., Briggs, A.W., Maricic, T., Stenzel, U., Kircher, M., Patterson, N., Li, H., Zhai, W., Fritz, M.H.-Y., et al. (2010). A draft sequence of the Neandertal genome. Science *328*, 710–722.

Groenen, M.A.M., Archibald, A.L., Uenishi, H., Tuggle, C.K., Takeuchi, Y., Rothschild, M.F., Rogel-Gaillard, C., Park, C., Milan, D., Megens, H.-J., et al. (2012). Analyses of pig genomes provide insight into porcine demography and evolution. Nature *491*, 393–398.

Harr, B., Karakoc, E., Neme, R., Teschke, M., Pfeifle, C., Pezer, Ž., Babiker, H., Linnenbrink, M., Montero, I., Scavetta, R., et al. (2016). Genomic resources for wild populations of the house mouse, *Mus musculus* and its close relative *Mus spretus*. Sci. Data *3*, 160075.

Havmøller, R.G., Payne, J., Ramono, W., Ellis, S., Yoganand, K., Long, B., Dinerstein, E., Williams, A.C., Putra, R.H., and Gawi, J. (2016). Will current conservation responses save the critically endangered Sumatran rhinoceros *Dicerorhinus sumatrensis*? Oryx *50*, 355–359.

Heintzman, P.D., Zazula, G.D., MacPhee, R., Scott, E., Cahill, J.A., McHorse, B.K., Kapp, J.D., Stiller, M., Wooller, M.J., Orlando, L., et al. (2017). A new genus of horse from Pleistocene North America. eLife *6*, e29944.

Heissig, K. (1972). Paläontologische und geologische Untersuchungen im Tertiär von Pakistan. V. Rhinocerotidae (Mamm.) aus den unteren und mittleren Siwalik-Schichten. Abh. Bayerischen Akad. Wiss. *152*, 1–112.

Heissig, K. (2012). The American genus *Penetrigonias* Tanner & Martin, 1976 (Mammalia: Rhinocerotidae) as a stem group elasmothere and ancestor of *Menoceras* Troxell, 1921 (Zitteliana).

Holt, C., and Yandell, M. (2011). MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. BMC Bioinformatics *12*, 491.

Junier, T., and Zdobnov, E.M. (2010). The Newick utilities: high-throughput phylogenetic tree processing in the UNIX shell. Bioinformatics *26*, 1669–1670.

Kahlke, R.-D., and Lacombat, F. (2008). The earliest immigration of woolly rhinoceros (*Coelodonta tologoijensis*, Rhinocerotidae, Mammalia) into Europe and its adaptive evolution in Palaearctic cold stage mammal faunas. Quat. Sci. Rev. *27*, 1951–1961.

Kalyaanamoorthy, S., Minh, B.Q., Wong, T.K.F., von Haeseler, A., and Jermiin, L.S. (2017). ModelFinder: fast model selection for accurate phylogenetic estimates. Nat. Methods *14*, 587–589.

Kardos, M., Åkesson, M., Fountain, T., Flagstad, Ø., Liberg, O., Olason, P., Sand, H., Wabakken, P., Wikenros, C., and Ellegren, H. (2018). Genomic consequences of intensive inbreeding in an isolated wolf population. Nat. Ecol. Evol. *2*, 124–131.

Karsten, M., Van Vuuren, B.J., Goodman, P.S., and Barnaud, A. (2011). The history and management of black rhino in KwaZulu-Natal: A population genetic approach to assess the past and guide the future. Anim. Conserv. *14*, 363–370.

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.

Khan, M.K.B.M., and van Strien, N.J. (1997). Asian rhinos: status survey and conservation action plan Volume 32 (IUCN).

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.

Kirillova, I.V., Chernova, O.F., Der Made, J.V., Kukarskih, V.V., Shapiro, B., Der Plicht, J.V., Shidlovskiy, F.K., Heintzmann, P.D., Van Kolfschoten, T., and Zanina, O.G. (2017). Discovery of the skull of *Stephanorhinus kirchbergensis* (Jäger, 1839) above the Arctic Circle. Quat. Res. *88*, 537–550.

Korf, I. (2004). Gene finding in novel genomes. BMC Bioinformatics *5*, 59, 59.

Korneliussen, T.S., Albrechtsen, A., and Nielsen, R. (2014). ANGSD: analysis of next generation sequencing data. BMC Bioinformatics *15*, 356.

Kosintsev, P., Mitchell, K.J., Devièse, T., van der Plicht, J., Kuitems, M., Petrova, E., Tikhonov, A., Higham, T., Comeskey, D., Turney, C., et al. (2019). Evolution and extinction of the giant rhinoceros *Elasmotherium sibiricum* sheds light on late Quaternary megafaunal extinctions. Nat. Ecol. Evol. *3*, 31–38.

Kosintsev, P.A., Zykov, S.V., Tiunov, M.P., Shpansky, A.V., Gasilin, V.V., Gimranov, D.O., and Devjashin, M.M. (2020). The first find of Merck's rhinoceros (Mammalia, Perissodactyla, Rhinocerotidae, *Stephanorhinus kirchbergensis* Jäger, 1839) remains in the Russian far east. Dokl. Biol. Sci. *491*, 47–49.

Kuhn, R.M., Haussler, D., and Kent, W.J. (2013). The UCSC genome browser and associated tools. Brief. Bioinform. *14*, 144–161.

Lewis, A.R., Marchant, D.R., Ashworth, A.C., Hedenas, L., Hemming, S.R., Johnson, J.V., Leng, M.J., Machlus, M.L., Newton, A.E., and Raine, J.I. (2008). Mid-Miocene cooling and the extinction of tundra in continental Antarctica. Proceedings of the National Academy of Sciences of the United States of America *105*, 10676–10680.

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.

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv. Genomics *1303*, 3997.

Li, H. (2014). Toward better understanding of artifacts in variant calling from high-coverage samples. Bioinformatics *30*, 2843–2851.

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics *25*, 1754–1760.

Li, H., and Durbin, R. (2011). Inference of human population history from individual whole-genome sequences. Nature *475*, 493–496.

Cell
Article

CellPress
OPEN ACCESS

Liu, S., Lorenzen, E.D., Fumagalli, M., Li, B., Harris, K., Xiong, Z., Zhou, L., Kor- neliussen, T.S., Somel, M., Babbitt, C., et al. (2014). Population genomics reveal recent speciation and rapid evolutionary adaptation in polar bears. Cell *157*, 785–794.

Lord, E., Dussex, N., Kierczak, M., Díez-Del-Molino, D., Ryder, O.A., Stanton, D.W.G., Gilbert, M.T.P., Sánchez-Barreiro, F., Zhang, G., Sinding, M.S., et al. (2020). Pre-extinction demographic stability and genomic signatures of adap- tation in the woolly rhinoceros. Curr. Biol. *30*, 3871–3879.e7.

Luo, R., Liu, B., Xie, Y., Li, Z., Huang, W., Yuan, J., He, G., Chen, Y., Pan, Q., Liu, Y., et al. (2015). Erratum: SOAPdenovo2: an empirically improved mem- ory-efficient short-read de novo assembler. Gigascience *4*, 30, 30.

Lynch, V.J., Bedoya-Reina, O.C., Ratan, A., Sulak, M., Drautz-Moses, D.I., Perry, G.H., Miller, W., and Schuster, S.C. (2015). Elephantid genomes reveal the molecular bases of woolly mammoth adaptations to the arctic. Cell Rep. *12*, 217–228.

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

Mallick, S., Li, H., Lipson, M., Mathieson, I., Gymrek, M., Racimo, F., Zhao, M., Chennagiri, N., Nordenfelt, S., Tandon, A., et al. (2016). The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. Nature *538*, 201–206.

Margaryan, A., Sinding, M.S., Liu, S., Vieira, F.G., Chan, Y.L., Nathan, S.K.S.S., Moodley, Y., Bruford, M.W., and Gilbert, M.T.P. (2020). Recent mitochondrial lineage extinction in the critically endangered Javan rhinoceros. Zool. J. Linn. Soc. *190*, 372–383.

Marshall, D.C., Simon, C., and Buckley, T.R. (2006). Accurate branch length estimation in partitioned Bayesian analyses requires accommodation of among-partition rate variation and attention to branch length priors. Syst. Biol. *55*, 993–1003.

Mays, H.L., Jr., Hung, C.M., Shaner, P.J.L., Denvir, J., Justice, M., Yang, S.F., Roth, T.L., Oehler, D.A., Fan, J., Rekulapally, S., and Primerano, D.A. (2018). Genomic analysis of demographic history and ecological niche modeling in the endangered Sumatran rhinoceros *Dicerorhinus sumatrensis*. Curr. Biol. *28*, 70–76.e4.

Mendes, F.K., and Hahn, M.W. (2016). Gene tree discordance causes apparent substitution rate variation. Syst. Biol. *65*, 711–721.

Meyer, M., and Kircher, M. (2010). Illumina sequencing library preparation for highly multiplexed target capture and sequencing. Cold Spring Harb. Protoc. *2010*, pdb.prot5448.

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.

Moodley, Y., Westbury, M.V., Russo, I.M., Gopalakrishnan, S., Rakotoarivelo, A., Olsen, R.-A., Prost, S., Tunstall, T., Ryder, O.A., Dalén, L., and Bruford, M.W. (2020). Interspecific gene flow and the evolution of specialisation in black and white rhinoceros. Mol. Biol. Evol. *37*, 3105–3117.

Muizon, C.d., and Ladevèze, S. (2020). Cranial anatomy of *Andinodelphys co- chabambensis*, a stem metatherian from the early Palaeocene of Bolivia. Geo- diversitas *42*, 597–739.

Nater, A., Mattle-Greminger, M.P., Nurcahyo, A., Nowak, M.G., de Manuel, M., Desai, T., Groves, C., Pybus, M., Sonay, T.B., Roos, C., et al. (2017). Morpho- metric, behavioral, and genomic evidence for a new orangutan species. Curr. Biol. *27*, 3576–3577.

Nguyen, L.T., Schmidt, H.A., von Haeseler, A., and Minh, B.Q. (2015). IQ- TREE: a fast and effective stochastic algorithm for estimating maximum-likeli- hood phylogenies. Mol. Biol. Evol. *32*, 268–274.

O'Leary, M.A., Bloch, J.I., Flynn, J.J., Gaudin, T.J., Giallombardo, A., Giannini, N.P., Goldberg, S.L., Kraatz, B.P., Luo, Z.X., Meng, J., et al. (2013). The placental mammal ancestor and the post-K-Pg radiation of placentals. Science *339*, 662–667.

Orlando, L., Ginolhac, A., Raghavan, M., Vilstrup, J., Rasmussen, M., Magnus- sen, K., Steinmann, K.E., Kapranov, P., Thompson, J.F., Zazula, G., et al. (2011). True single-molecule DNA sequencing of a pleistocene horse bone. Genome Res. *21*, 1705–1719.

Orlando, L., Leonard, J.A., Thenot, A., Laudet, V., Guérin, C., and Hänni, C. (2003). Ancient DNA analysis reveals woolly rhino evolutionary relationships. Mol. Phylogenet. Evol. *28*, 485–499.

Owen-Smith, N. (2015). Mechanisms of coexistence in diverse herbivore– carnivore assemblages: demographic, temporal and spatial heterogeneities affecting prey vulnerability. Oikos *124*, 1417–1426.

Pääbo, S. (1989). Ancient DNA: extraction, characterization, molecular clon- ing, and enzymatic amplification. Proc. Natl. Acad. Sci. USA *86*, 1939–1943.

Palkopoulou, E., Mallick, S., Skoglund, P., Enk, J., Rohland, N., Li, H., Omrak, A., Vartanyan, S., Poinar, H., Götherström, A., et al. (2015). Complete genomes reveal signatures of demographic and genetic declines in the woolly mammoth. Curr. Biol. *25*, 1395–1400.

Palkopoulou, E., Lipson, M., Mallick, S., Nielsen, S., Rohland, N., Baleka, S., Karpinski, E., Ivancevic, A.M., To, T.H., Kortschak, R.D., et al. (2018). A comprehensive genomic history of extinct and living elephants. Proc. Natl. Acad. Sci. USA *115*, E2566–E2574.

Parks, M., and Lambert, D. (2015). Impacts of low coverage depths and post- mortem DNA damage on variant calling: a simulation study. BMC Genomics *16*, 19, 19.

Pickford, M., Senut, B., and Hadoto, D. (1993). Geology and palaeobiology of the Albertine Rift valley, Uganda-Zaire. Volume I: geology; Géologie et paléo- biologie de la vallée du Rift albertin, Ouganda, Zaïre. Volume I: Géologie (CI- FEG Occasional Publications).

Player, I. (1973). The White Rhino Saga, 1st (Stein and Day).

Prado-Martinez, J., Sudmant, P.H., Kidd, J.M., Li, H., Kelley, J.L., Lorente-Gal- dos, B., Veeramah, K.R., Woerner, A.E., O'Connor, T.D., Santpere, G., et al. (2013). Great ape genetic diversity and population history. Nature *499*, 471–475.

Prüfer, K., Munch, K., Hellmann, I., Akagi, K., Miller, J.R., Walenz, B., Koren, S., Sutton, G., Kodira, C., Winer, R., et al. (2012). The bonobo genome compared with the chimpanzee and human genomes. Nature *486*, 527–531.

Reddy, P.C., Sinha, I., Kelkar, A., Habib, F., Pradhan, S.J., Sukumar, R., and Galande, S. (2015). Comparative sequence analyses of genome and transcrip- tome reveal novel transcripts and variants in the Asian elephant *Elephas max- imus*. J. Biosci. *40*, 891–907.

Robinson, J.A., Ortega-Del Vecchyo, D., Fan, Z., Kim, B.Y., vonHoldt, B.M., Marsden, C.D., Lohmueller, K.E., and Wayne, R.K. (2016). Genomic flatlining in the endangered island fox. Curr. Biol. *26*, 1183–1189.

Rookmaaker, K., and Antoine, P.-O. (2012). New maps representing the histor- ical and recent distribution of the African species of rhinoceros: *Diceros bicor- nis, Ceratotherium simum* and *Ceratotherium cottoni*. Pachyderm *52*, 91–96.

Rose, K.D., Holbrook, L.T., Rana, R.S., Kumar, K., Jones, K.E., Ahrens, H.E., Missiaen, P., Sahni, A., and Smith, T. (2014). Early Eocene fossils suggest that the mammalian order Perissodactyla originated in India. Nat. Commun. *5*, 5570.

Sánchez-Barreiro, F., Gopalakrishnan, S., Ramos-Madrigal, J., Westbury, M.V., de Manuel, M., Margaryan, A., Ciucani, M.M., Vieira, F.G., Patramanis, Y., Kalthoff, D.C., et al. (2021). Historical population declines prompted signif- icant genomic erosion in the northern and southern white rhinoceros. Mol. Ecol. Published online June 27, 2021. https://doi.org/10.1111/mec.16043.

Sayyari, E., Whitfield, J.B., and Mirarab, S. (2018). DiscoVista: Interpretable vi- sualizations of gene tree discordance. Mol. Phylogenet. Evol. *122*, 110–115.

Schubert, M., Lindgreen, S., and Orlando, L. (2016). AdapterRemoval v2: rapid adapter trimming, identification, and read merging. BMC Res. Notes *9*, 88.

Seppey, M., Manni, M., and Zdobnov, E.M. (2019). BUSCO: Assessing Genome Assembly and Annotation Completeness. Methods Mol. Biol. *1962*, 227–245.

Shpansky, A.V., and Boeskorov, G.G. (2018). Northernmost Record of the Merck's Rhinoceros *Stephanorhinus kirchbergensis* (Jäger) and Taxonomic

Status of *Coelodonta jacuticus* Russanov (Mammalia, Rhinocerotidae). Pale-ontol. J. *52*, 445–462.

Simpson, J.T., Wong, K., Jackman, S.D., Schein, J.E., Jones, S.J.M., and Birol, I. (2009). ABySS: a parallel assembler for short read sequence data. Genome Res. *19*, 1117–1123.

Slater, G.S.C., and Birney, E. (2005). Automated generation of heuristics for biological sequence comparison. BMC Bioinformatics *6*, 31.

Smith, C.D., Edgar, R.C., Yandell, M.D., Smith, D.R., Celniker, S.E., Myers, E.W., and Karpen, G.H. (2007). Improved repeat identification and masking in Dipterans. Gene *389*, 1–9.

Smith, S.D., Kawash, J.K., Karaiskos, S., Biluck, I., and Grigoriev, A. (2017). Evolutionary adaptation revealed by comparative genome analysis of woolly mammoths and elephants. DNA Res. *24*, 359–369.

Sosdian, S.M., Babila, T.L., Greenop, R., Foster, G.L., and Lear, C.H. (2020). Ocean Carbon Storage across the middle Miocene: a new interpretation for the Monterey Event. Nat. Commun. *11*, 134.

Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics *30*, 1312–1313.

Stanke, M., Diekhans, M., Baertsch, R., and Haussler, D. (2008). Using native and syntenically mapped cDNA alignments to improve de novo gene finding. Bioinformatics *24*, 637–644.

Steiner, C.C., and Ryder, O.A. (2011). Molecular phylogeny and evolution of the Perissodactyla. Zool. J. Linn. Soc. *163*, 1289–1303.

Suyama, M., Torrents, D., and Bork, P. (2006). PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. Nu-cleic Acids Res. *34*, W609-12.

Svardal, H., Jasinska, A.J., Apetrei, C., Coppola, G., Huang, Y., Schmitt, C.A., Jacquelin, B., Ramensky, V., Müller-Trutwin, M., Antonio, M., et al. (2017). Ancient hybridization and strong adaptation to viruses across African vervet monkey populations. Nat. Genet. *49*, 1705–1713.

Teng, H., Zhang, Y., Shi, C., Mao, F., Cai, W., Lu, L., Zhao, F., Sun, Z., and Zhang, J. (2017). Population genomics reveals speciation and introgression between brown Norway rats and their sibling species. Mol. Biol. Evol. *34*, 2214–2228.

Thorne, J.L., Kishino, H., and Painter, I.S. (1998). Estimating the rate of evolu-tion of the rate of molecular evolution. Mol. Biol. Evol. *15*, 1647–1657.

Tougard, C., Delefosse, T., Hänni, C., and Montgelard, C. (2001). Phylogenetic relationships of the five extant Rhinoceros species (Rhinocerotidae, Perisso-dactyla) based on mitochondrial cytochrome b and 12S rRNA genes. Mol. Phy-logenet. Evol. *19*, 34–44.

Tunstall, T., Kock, R., Vahala, J., Diekhans, M., Fiddes, I., Armstrong, J., Paten, B., Ryder, O.A., and Steiner, C.C. (2018). Evaluating recovery potential of the northern white rhinoceros from cryopreserved somatic cells. Genome Res. *28*, 780–788.

Van Couvering, J.A., and Delson, E. (2020). African land mammal ages. J. Vertebr. Paleontol. *40*, e1803340.

Vandenberghe, N., Hilgen, F.J., and Speijer, R.P. (2012). The Paleogene period. In The Geologic Time Scale 2012, F.M. Gradstein, J.G. Ogg, M.D. Schmitz, and G.M. Ogg, eds. (Elsevier Science), pp. 855–921.

Veeramah, K.R., Woerner, A.E., Johnstone, L., Gut, I., Gut, M., Marques-Bonet, T., Carbone, L., Wall, J.D., and Hammer, M.F. (2015). Examining phylo-genetic relationships among gibbon genera using whole genome sequence data using an approximate bayesian computation approach. Genetics *200*, 295–308.

von Seth, J., Dussex, N., Díez-Del-Molino, D., van der Valk, T., Kutschera, V.E., Kierczak, M., Steiner, C.C., Liu, S., Gilbert, M.T.P., Sinding, M.-S., et al. (2021). Genomic insights into the conservation status of the world's last remaining Su-matran rhinoceros populations. Nat. Commun. *12*, 2393.

Wang, K., Lenstra, J.A., Liu, L., Hu, Q., Ma, T., Qiu, Q., and Liu, J. (2018). Incomplete lineage sorting rather than hybridization explains the inconsistent phylogeny of the wisent. Commun. Biol. *1*, 169.

Wang, G.-D., Zhang, M., Wang, X., Yang, M.A., Cao, P., Liu, F., Lu, H., Feng, X., Skoglund, P., Wang, L., et al. (2019). Genomic approaches reveal an endemic subpopulation of gray wolves in southern China. iScience *20*, 110–118.

Welker, F., Collins, M.J., Thomas, J.A., Wadsley, M., Brace, S., Cappellini, E., Turvey, S.T., Reguero, M., Gelfo, J.N., Kramarz, A., et al. (2015). Ancient pro-teins resolve the evolutionary history of Darwin's South American ungulates. Nature *522*, 81–84.

Welker, F., Smith, G.M., Hutson, J.M., Kindler, L., Garcia-Moreno, A., Villa-luenga, A., Turner, E., and Gaudzinski-Windheuser, S. (2017). Middle Pleisto-cene protein sequences from the rhinoceros genus *Stephanorhinus* and the phylogeny of extant and extinct Middle/Late Pleistocene Rhinocerotidae. PeerJ *5*, e3033.

Westbury, M., Baleka, S., Barlow, A., Hartmann, S., Paijmans, J.L.A., Kramarz, A., Forasiepi, A.M., Bond, M., Gelfo, J.N., Reguero, M.A., et al. (2017). A mito-genomic timetree for Darwin's enigmatic South American mammal *Macrau-chenia patachonica*. Nat. Commun. *8*, 15951.

Westbury, M.V., Hartmann, S., Barlow, A., Wiesel, I., Leo, V., Welch, R., Parker, D.M., Sicks, F., Ludwig, A., Dalén, L., and Hofreiter, M. (2018). Extended and continuous decline in effective population size results in low genomic diversity in the world's rarest hyena species, the brown hyena. Mol. Biol. Evol. *35*, 1225–1237.

Westbury, M.V., Petersen, B., Garde, E., Heide-Jørgensen, M.P., and Loren-zen, E.D. (2019). Narwhal genome reveals long-term low genetic diversity despite current large abundance size. iScience *15*, 592–599.

Wilkie, G.S., Davison, A.J., Watson, M., Kerr, K., Sanderson, S., Bouts, T., Steinbach, F., and Dastjerdi, A. (2013). Complete genome sequences of elephant endotheliotropic herpesviruses 1A and 1B determined directly from fatal cases. J. Virol. *87*, 6700–6712.

Willerslev, E., Gilbert, M.T.P., Binladen, J., Ho, S.Y., Campos, P.F., Ratan, A., Tomsho, L.P., da Fonseca, R.R., Sher, A., Kuznetsova, T.V., et al. (2009). Anal-ysis of complete mitochondrial genomes from extinct and extant rhinoceroses reveals lack of phylogenetic resolution. BMC Evol. Biol. *9*, 95.

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). Moun-tain gorilla genomes reveal the impact of long-term population decline and inbreeding. Science *348*, 242–245.

Yang, Z. (2007). PAML 4: phylogenetic analysis by maximum likelihood. Mol. Biol. Evol. *24*, 1586–1591.

Yang, D.Y., Eng, B., Waye, J.S., Dudar, J.C., and Saunders, S.R. (1998). Tech-nical note: improved DNA extraction from ancient bones using silica-based spin columns. Am. J. Phys. Anthropol. *105*, 539–543.

Yates, A.D., Achuthan, P., Akanni, W., Allen, J., Allen, J., Alvarez-Jarreta, J., Amode, M.R., Armean, I.M., Azov, A.G., Bennett, R., et al. (2020). Ensembl 2020. Nucleic Acids Res. *48* (D1), D682–D688.

Yim, H.-S., Cho, Y.S., Guang, X., Kang, S.G., Jeong, J.-Y., Cha, S.-S., Oh, H.-M., Lee, J.-H., Yang, E.C., Kwon, K.K., et al. (2014). Minke whale genome and aquatic adaptation in cetaceans. Nat. Genet. *46*, 88–92.

Yoo, D., Kim, K., Kim, H., Cho, S., Kim, J.N., Lim, D., Choi, S.-G., Choi, B.-H., and Kim, H. (2017). The genetic origin of short tail in endangered Korean dog, DongGyeongi. Sci. Rep. *7*, 10048.

Zhang, W., Fan, Z., Han, E., Hou, R., Zhang, L., Galaverni, M., Huang, J., Liu, H., Silva, P., Li, P., et al. (2014). Hypoxia adaptations in the grey wolf (*Canis lupus* chanco) from Qinghai-Tibet Plateau. PLoS Genet. *10*, e1004466.

Zheng, Y., and Janke, A. (2018). Gene flow analysis method, the D-statistic, is robust in a wide parameter space. BMC Bioinformatics *19*, 10.

Zoonomia Consortium (2020). A comparative genomics multitool for scientific discovery and conservation. Nature *587*, 240–245.

Cell
Article

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Biological samples		
Javan rhinoceros	This study	NHM, UiO 734; NCBI:txid102233
Merck's rhinoceros	This study	F-4160; NCBI:txid2003782
Siberian Unicorn	This study	IPAE 915/2804; NCBI:txid2491732
Woolly rhinoceros	This study	ND036; NCBI:txid222863
Greater one-horned rhinoceros	This study	KB14498/SB137; NCBI:txid9809
Black rhinoceros	Moodley et al., 2020	GCA_013634535.1; NCBI:txid9805
Sumatran rhinoceros	Lord et al., 2020	GCA_014189135.1; NCBI:txid89632
White rhinoceros	SAMN00778988	GCF_000283155.1; NCBI:txid73337
Deposited data		
Raw sequence data and assemblies	This study	NCBI Project Number PRJNA687817
Custom scripts for genome alignment analysis	This study	https://github.com/liushanlin/rhinoceros-comparative-genome
Software and algorithms		
AdapterRemoval v2.2.2	Schubert et al., 2016	https://github.com/MikkelSchubert/adapterremoval
ALLPATHS-LG v.52485	Gnerre et al., 2011	http://software.broadinstitute.org/allpaths-lg/blog/?page_id=12
ABySS v.2.0.2	Simpson et al., 2009	https://github.com/bcgsc/abyss
SOAPdenovo2	Luo et al., 2015	https://github.com/aquaskyline/SOAPdenovo2
BUSCO v.5.0.0	Seppey et al., 2019	https://busco.ezlab.org/
RepeatMasker Open-4.0.2013-2015	Arian Smit & Robert Hubley	http://www.repeatmasker.org/
RepeatRunner	Smith et al., 2007	https://www.yandell-lab.org/software/repeatrunner.html
Augustus v3.2.3	Stanke et al., 2008	http://bioinf.uni-greifswald.de/augustus/
SNAP	Korf, 2004	https://github.com/KorfLab/SNAP
Maker v2.31.10	Holt and Yandell, 2011	https://www.yandell-lab.org/software/maker.html
BWA v0.7.17-r1188	Li and Durbin, 2009	https://github.com/lh3/bwa
LAST v1061	Kiełbasa et al., 2011	https://gitlab.com/mcfrith/last
MULTIZ v11.2	Blanchette et al., 2004	http://www.bx.psu.edu/miller_lab/
ANGSD v0.924)	Korneliussen et al., 2014	http://www.popgen.dk/angsd/index.php/ANGSD
MAFFT v7.310	Katoh and Standley, 2013	https://mafft.cbrc.jp/alignment/software/linuxportable.html
UCSC tools	Kuhn et al., 2013	https://github.com/ucscGenomeBrowser/kent
RaxML v8.2.10	Stamatakis, 2014	https://github.com/stamatak/standard-RAxML
ASTRAL v5.6.3	Mirarab et al., 2014	https://github.com/smirarab/ASTRAL
DiscoVista v1.0	Sayyari et al., 2018	https://github.com/esayyari/DiscoVista
NW_utiles v1.5.0	Junier and Zdobnov, 2010	https://github.com/tjunier/newick_utils/wiki
EXONERATE v2.2.0	Slater and Birney, 2005	https://www.ebi.ac.uk/about/vertebrate-genomics/software/exonerate
PAL2NAL v14	Suyama et al., 2006	https://github.com/drostlab/orthologr/tree/master/inst/pal2nal/pal2nal.v14
MCMCTree v4.8	Yang, 2007	http://abacus.gene.ucl.ac.uk/software/paml.html#download

(Continued on next page)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
PSMC	Li and Durbin, 2011	https://github.com/lh3/psmc
SnpEff	Cingolani et al., 2012	https://github.com/pcingola/SnpEff
Bcftools v1.8	Li, 2011	https://samtools.github.io/bcftools/
GATK v3.8	Auwera et al., 2013	https://gatk.broadinstitute.org/hc/
IQ-Tree v1.6	Nguyen et al., 2015	https://github.com/Cibiv/IQ-TREE

RESOURCE AVAILABILITY

Lead contact
Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact: Shanlin Liu (shanlin.liu@cau.edu.cn).

Materials availability
This study did not generate any new unique reagents.

Data and code availability
Raw sequencing data and genome assemblies can be accessed at NCBI databases under project number Bio-Project:PRJNA687817. The custom scripts for genome alignment analysis for the ancient and historical rhinoceros have been deposited in https://github.com/liushanlin/rhinoceros-comparative-genome.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Source organisms
The specimen of the greater one-horned [Indian] rhinoceros (*Rhinoceros unicornis*) derives from cell culture held at the San Diego Zoo (ID = KB14498, SB137) and originally derives from a captive born female individual, that in turn was derived from captive born parents (Dam-100288 and Sire-100289). The specimen of black rhinoceros (*Diceros bicornis*) was provided by the Zululand Rhino Reserve, South Africa (ID = 46373), and is detailed further in the original paper reporting its genome (Moodley et al., 2020). The Sumatran rhinoceros (*Dicerorhinus sumatrensis*) sample derives from blood sampled from an individual named Kertam by a staff at the Sabah Wildlife Department in Borneo, Malaysia. Further details on this specimen are detailed in the original paper that released its genome (Lord et al., 2020). The Javan rhinoceros (*Rhinoceros sondaicus*) genome derives from three sub-samples of dried soft tissue (two) and bone (one) taken from a skull collected in 1838 in Java, that is currently kept in the collections of the Natural History Museum at the University of Oslo (Natural History Museum, Oso; accession Museum id: 734). The woolly rhinoceros (*Coelodonta antiquitatis*) specimen (ID = ND036; a femur) was collected along the Rakvachan River (N69° 17,80' E167° 38,53'), on the Kyttyk Peninsula, Chukotka, Russia. The sample has been radiocarbon dated twice, yielding radiocarbon age estimates of 46200 ± 2300 ^{14}C years BP (OxA-36569) and 51980 ± 4900 14C years BP (MAG-2095). The Siberian unicorn (*Elasmotherium sibiricum*) genome was generated from a radius subsample of a specimen (IPAE 915/2804) that originated from Tobolsk, Russia (58°N 68°E) and has a radiocarbon age of > 49,200 ^{14}C years BP (OxA-34900). The genome of the Merck's rhinoceros (*Stephanorhinus kirchbergensis*) derives from a first molar (M1) tooth root that was subsampled from a complete cranium (National Alliance of Shidlovskiy "Ice Age," Ice Age Museum, Moscow; accession F-4160) recovered from the Chondon River valley in Yakutia, Russia (N70° 12′ E137°) and dated to between 48,000 and 70,000 years BP (Kirillova et al., 2017; Cappellini et al., 2019). Lastly the white rhinoceros (*Ceratotherium simum simum*) genome was taken from the Broad Institute assembly that is publicly available at https://www.ncbi.nlm.nih.gov/assembly/GCF_000283155.1/.

METHOD DETAILS

DNA isolation and sequencing
DNA of the Javan rhinoceros was extracted and prepared for sequencing in the ancient DNA laboratories at the GLOBE Institute, University of Copenhagen, following standard clean lab procedures (Cappellini et al., 2012; Orlando et al., 2011). DNA was extracted from dried soft tissue using a SDS - DTT - proteinase k buffer (Gilbert et al., 2007) and from bone using a EDTA - urea - proteinase k buffer (Ersmark et al., 2015). Libraries were constructed following the BEST protocol (Caroe et al., 2017) with the modifications as in Mak et al. (2017), and sequenced on the BGISeq500 platform at BGI Shenzhen, with a sequencing strategy of 50 PE.

DNA was extracted from the woolly rhinoceros specimen at the ancient DNA laboratory at the Swedish Museum of Natural History from 50 mg of bone powder collected using a Dremel drill. Then, we extracted DNA using a modified version of protocol C in Yang et al. (1998) as described in Brace et al. (2012). Sequencing libraries were prepared following the BEST library build protocol (Caroe et al., 2017), where the adaptor oligos were custom-designed for the BGISeq 500 Sequencing Platform (Mak et al., 2017). Libraries were amplified in a 50 μL reaction containing 5 U AmpliTaq Gold polymerase (Applied Biosystems, Foster City, CA), 1× AmpliTaqGold buffer, 2.5 mM MgCl2, 0.4 mg/mL bovine serum albumin (BSA), 0.2 mM each dNTP, 0.2 μM BGI forward primer (Mak et al., 2017), 0.2 μM BGI reverse index-primer (Mak et al., 2017) and 10 μL of library DNA template. Amplified libraries were sequenced on a BGISeq 500 platform at BGI Shenzhen with a SE100 sequencing strategy.

We extracted DNA from the Merck's rhinoceros M1 tooth root, as previously reported by Kirillova et al. (2017) and generated data as described in Cappellini et al. (2019). Briefly, in addition to the DNA extract obtained in Kirillova et al., a second DNA extraction was performed following the previously reported methods (Dabney et al., 2013) in the University of California Santa Cruz Paleogenomics laboratory. We constructed four libraries (single-indexed and double-stranded) for each DNA extract following the methods reported by Meyer and Kircher (2010). The total of eight indexed libraries were pooled and sent to SciLifeLab (Stockholm, Sweden) for sequencing on two lanes of the Illumina HiSeq-X platform (150 PE). In addition, a subsample was processed in the ancient DNA laboratories at the GLOBE Institute, University of Copenhagen, following the same protocols as described with the same library construction method for the Javan rhinoceros, and sequenced on three lanes of PE100 data on a BGISeq 500 platform at BGI Shenzhen.

A subsample was taken from a radius (IPAE 915/2804) of the Siberian unicorn and DNA was extracted at the Australian Centre for Ancient DNA (University of Adelaide) as previously reported by Kosintsev et al. (2019). A new Illumina sequencing library was created for this study following the protocol described by Meyer and Kircher (2010) with an additional step added to excise deaminated cytosines with the USER enzyme mix (New England Biolabs) as described by Briggs et al., (2010). The library was split into 8 separate PCR reactions to minimize PCR bias and maintain library complexity. Each PCR of 25 μL contained 1 × HiFi buffer, 2.5 mM MgSO4, 1 mM dNTPs, 0.5 mM each primer (containing a unique combination of 7-mer i5 and i7 indexes), 0.1 U Platinum Taq Hi-Fi polymerase and 3 μL DNA. The cycling conditions were 94°C for 6 min; 7 cycles of 94°C for 30 s, 60°C for 30 s, and 68°C for 40 s; followed by 68°C for 10 min. Following PCR, replicates were pooled and purified using 1.1x volume AxyPrep magnetic beads, eluted in 30 μL EB buffer, and quantified using a TapeStation (Agilent Technologies). The library was sequenced on three lanes of an Illumina HiSeq X Ten at the Garvan Institute of Medical Research (Sydney, Australia). 4 additional PCR reactions were conducted in Stockholm, Sweden. Each PCR of 25 μL contained 1x AccuPrime reaction mix, 0.3 μM IS4 amplification primer, 0.3 μM P7 indexing primer, 7 U AccuPrime Pfx (Thermo Scientific) polymerase. The cycling conditions were 95°C for 2 min, 14 cycles at 95°C for 15 s, 60°C for 30 s and 68°C for 1 min. PCR replicates were pooled and purified using Agencourt AMPure XP beads (Beckman Coulter, Brea, CA, USA), eluted in 36 μL EB Buffer, and quantified using a high-sensitivity DNA chip on a Bioanalyzer 2100 (Agilent, Santa Clara, CA, USA). The library was subsequently sequenced on one Illumina HiSeqX lane with a 2 × 151 bp setup in the High Output mode at SciLifeLab (Stockholm, Sweden).

For the historic/ancient rhinoceros samples, although multiple sequencing libraries were constructed and some were sequenced with different read length strategies, comparison tests of the different datasets found no bias differences (detailed in the Additional resources section) and thus we merged them for the subsequent analyses.

High molecular weight DNA was extracted from a cell culture specimen from the San Diego Zoo (ID = KB14498, SB137) for the greater one-horned rhinoceros using a Kingfisher duo prime extraction robot. Paired-end Truseq PCR-free libraries of insert size of 180bp and 670bp were constructed and sequenced on the Illumina HiSeq X platform at SciLifeLab (Stockholm, Sweden) generating approximately 400 million paired-end (2x150bp) each. Additionally, three mate-pair libraries (3, 5 and 20 kb) from two specimens, one from San Diego Zoo (KB17733) and one from Rotterdam Zoo (IR_104724), were constructed and sequenced for approximately 100 million paired-end reads (2x150bp) each on the Illumina HiSeq X at SciLifeLab (Stockholm, Sweden).

Genome assembly and annotation

We merged the paired-end reads for all the ancient and historic rhino samples (Siberian unicorn, Javan, Merck's and woolly rhinoceros) using a modified version of AdapterRemoval (Schubert et al., 2016) which masked the conflict bases that have identical sequencing qualities to Ns and removed collapsed reads < 30 bp. Then, we mapped the short reads from those rhinoceros species onto their corresponding genome references (Table S2) and obtained their genomes (consensus sequences) using the doFasta module (-doFasta 2) in ANGSD (version 0.924) (Korneliussen et al., 2014) with a minimum requirement of mapping quality and base quality of 20. A minimum depth of 5 was set for the Javan and woolly rhinoceros, and the minimum depth was set as 3 for the Siberian unicorn and Merck's rhinoceros due to their relatively lower coverages (~30× for the former pair, versus ~10× for the latter pair).

For the greater one-horned rhinoceros, we obtained its genome using a mate-pair assembly. Since assembly quality can vary for different datasets, we used three different assemblers and evaluated their performance. The following assemblers for short read sequence data were used: ALLPATHS-LG v.52485 (HAPLOIDIFY = True) (Gnerre et al., 2011), ABySS v.2.0.2 (-k = 61) (Simpson et al., 2009) and SOAPdenovo2 v. 2.04 (-K 61) (Luo et al., 2015). Out of the three assemblers, ALLPATHS-LG was selected for downstream analyses as it produced the most gene-complete and the contiguous assembly (with a scaffold N50 of 27.7 Mbp. Gene completeness was measured with BUSCO v.5.0.0 (Seppey et al., 2019) using the "mammalia_odb10" ortholog dataset, which showed a low degree of missing, fragmented and duplicated genes: "C:96.2% [S:95.6%, D:0.6%], F:1.0%, M:2.8%, n:9226."

The genome annotation of white rhinoceros is its original release with its genome assembly. Since genome annotations of both the black and Sumatran rhinoceros were not released with their genome assemblies, we annotated the genomes of the black, Sumatran and greater one-horned rhinoceroses as follows. First, repeats were masked using RepeatMasker (Smit, AFA, Hubley, R & Green, P. RepeatMasker Open-4.0.2013-2015) with all ("model_org=all") species being included in the RepBase. Then, we masked the transposable element proteins using RepeatRunner (Smith et al., 2007) and the repeat protein library te_proteins.fasta (downloaded from http://weatherby.genetics.utah.edu/data/maker_tutorial.tgz). After that, we applied Ensemble (version 64) (Yates et al., 2020) for homolog prediction with its amino acid datasets including *Sus scrofa*, *Tursiops truncatus*, *Bos taurus*, *Drosophila melanogaster*, *Vicugna pacos* and *Homo sapiens*. *De novo* gene prediction was achieved with: a) Augustus (version 3.2.3) (Stanke et al., 2008), where "human" was used as the gene prediction species model for Augustus; b) and SNAP (Korf, 2004) using the provided mammal model (mam54). Finally, Maker (version 2.31.10) (Holt and Yandell, 2011) was used to weigh and merge different evidence, and obtain protein-coding-gene sets for the black, Sumatran and greater one-horned rhinoceros.

Genome alignment and phylogenetic inference

The four *de novo* assembled rhinoceros genomes and the genome of tapir (*Tapirus terrestris*, GCA_004025025.1) were aligned to the horse reference genome (*Equus caballus*, GCF_000002305.2) using LAST with parameters of -m 100 -E 0.05 (Kiełbasa et al., 2011), and each pairwise alignment was chained and netted to form high quality blocks. After that, the non-syntenic regions were filtered out from each alignment. Finally, the pairwise local alignments were combined to generate a final multi-species whole genome alignment (MWGA) using MULTIZ without reference fixed and with a radius of 50 in dynamic programming (Blanchette et al., 2004).

After generating the MWGA for all the *de novo* assemblies, a sliding window-based phylogenetic analysis was conducted along the horse genome with a window size of 100 kb. A set of UCSC genome browser tools (Kuhn et al., 2013), including mafsInRegion, mafFilter, maf2fasta, and several in-house PERL scripts were applied to divide the whole genome alignment into sub-regions. Then, for each sub-region, we included the genome sequences of the non-*de novo* sequenced rhinoceros samples, via extracting the corresponding sequences based on the region coordinate information on their reference genomes, and conducted multiple sequence alignment using MAFFT (Katoh and Standley, 2013). We also removed the alignments with effective length < 1,000 bp and effective ratio < 0.5, of which the effective sites represent nucleotide sites that do not include missing information (Ns) for any taxa in the multiple sequence alignment. After that, the GTR+CAT model of RaxML (Stamatakis, 2014) was used to build Maximum Likelihood (ML) phylogenetic trees for each window. Finally, the species tree was generated using the multi-species coalescent model-based software ASTRAL III (Mirarab et al., 2014), by combining all the regional trees.

Due to the poor quality of the endogenous DNA recovered from the historical and ancient samples, their genome sequences could only be reconstructed through mapping against a reference genome. To test whether the choice of reference genome would play any role in shaping either the consensus sequence recovered, or phylogenetic inference based on the data, we used the 100 PE reads generated from the woolly rhinoceros and the Siberian unicorn to examine reference bias via aligning their short reads onto different references including the horse (*Equus caballus*), white, greater one-horned, and Sumatran rhinoceros, respectively. Following genome alignment, 82 genome regions of length > 20,000 bp were selected to infer phylogenetic trees (Figure S1A). Those regions were randomly scattered across the genome, thus having little chance of being found in the same recombination blocks.

We calculated the frequency of three topologies around each focal internal branch of the species tree (Figure 2B) using DiscoVista (Sayyari et al., 2018) and NW_utils (Junier and Zdobnov, 2010), for the 100 kb windows based gene trees. A multi-species coalescent simulation was also applied to determine the expected gene tree distributions on the basis of the dated species tree using the Phybase package following Wang et al. (2018). Then, we inspected the congruency between the frequency of the three topologies inferred from empirical genomic data and that generated from the simulation for the Rhinocerotinae lineage.

To validate the robustness of our species tree, we also reconstructed the phylogenetic relationship across each chromosome independently (reference by horse's chromosomes) for all the rhinoceros species that have *de novo* genome assemblies and two outgroups of tapir and horse using the aforementioned 100 kb sliding windows based method. Then, for each chromosome we inferred a species tree and calculated the tree topology frequency using DiscoVista (Sayyari et al., 2018).

As a genomic region with length of 100 kb may contain multiple recombination breakpoints, we sampled a short alignment with length of 5,000 bp within each 100 kb sliding window to infer gene trees using RaxML (Stamatakis, 2014) with a substitution model setting of GTR+CAT and 100 bootstrap replicates. Then we filtered out gene trees with nodes of bootstrap support < 85 to guarantee congruent signals throughout each subregion. Finally, the frequency of three topologies around the phylogenetic discordance branch obversed based on 100 kb sliding windows were calculated using DiscoVista (Sayyari et al., 2018).

Molecular dating

Orthologs of the species with *de novo* genome assemblies were extracted from the former synteny checked MWGA: 1) gene location and the corresponding exon regions in the horse genome annotation file were used to extract CDSs from the MWGA alignment asking for a coverage $\geq$ 80%; 2) the amino acid sequences obtained from the horse genome serving as query were used to find corresponding homologs for each species using protein2genome model in EXONERATE (Slater and Birney, 2005); 3) exons that were not shared by all species were removed to improve the gene alignment accuracy; 4) amino acid sequences were aligned using MAFFT (Katoh and Standley, 2013) and in turn used to guide the CDS alignment using PAL2NAL (Suyama et al., 2006). For the samples without *de novo* assemblies, we obtained their orthologs according to the gene and exon location information from their reference genomes. In

Cell
Article

order to further diminish the potential influence of reference bias, we obtained two gene sequences for each ortholog by mapping their reads onto two different reference genomes (Table S2), and merged the two gene sequences while masking conflicts as "N" (X for amino acids).

We inferred the evolutionary timescale of rhinoceros lineages using a set of 3,513 orthologs from the set identified in the previous section (8,820,642 nucleotides). The set of orthologs was selected to minimize the possible bias in molecular rates arising from analyses of finite number of sites or excessive discrepancies in molecular rates across lineages in particular loci (Marshall et al., 2006). To select orthologs, we first performed a fast tree search for each ortholog, using only NNI tree rearrangements, under a GTR+R4 substitution model (Kalyaanamoorthy et al., 2017) implemented in IQ-TREE v1.6 (Nguyen et al., 2015). Orthologs were considered as adequate for molecular dating analyses if they contained low degrees of rate variation among lineages (Doyle et al., 2015), and if their gene trees were not too greatly distinct from the species tree (Mendes and Hahn 2016). Specifically, the orthologs retained had gene trees with coefficients of variation in root-to-tip length (non-clocklikeness) < 0.1, and Robinson-Foulds distances to our inferred species tree ≤ 2.

Our analyses included time-calibrations on four splits of the species tree, each based on several lines of evidence from the fossil record. The nodes were calibrated using uniform distributions with soft maximum bounds, placing a 0.025 probability on older ages. Placement of maximum bound for a clade assumes that less inclusive clades cannot be older than the oldest fossils of more inclusive clades. (i) We calibrated the split between Elasmotheriinae and Rhinocerotinae as occurring between 35 Ma and 44 Ma. This minimum bound is supported by the earliest record of Rhinocerotinae, with *Epiaceratherium naduongense*, at between 35-39 Ma (Böhme et al., 2013), which is unambiguously a species nested within Rhinocerotinae. Additional evidence for this minimum bound is the earliest species assigned to Elasmotheriinae, *Penetrigonas dakotensis*, which appears at ca. 38 Ma (Heissig, 2012). Furthermore, morpho-anatomical cladistic analysis has placed the 38 My-old fossil *Subhyracodon occidentali*s as a representative of Elasmotheriinae (Becker et al., 2013). The maximum bound is supported by fossil and phylogenetic evidence placing with high confidence the minimum age of Rhinocerotidae at 44 Ma (Bai et al., 2020). (ii) The split between *Rhinoceros* and *Dicerorhinus* was calibrated as occurring between 13 Ma and 23 Ma. The minimum age is based on remains from the middle Siwaliks of Pakistan dating from ca. 13 Ma and unambiguously assigned to *Dicerorhinus*, under the name of *D.* aff. *sumatrensis* (Heissig, 1972). Other remains from the same region and age confirm the occurrence of *Dicerorhinus* as early as the middle Miocene (13.7-11.65 Ma; P.-O.A., unpublished data). The maximum bound is informed by the first attested occurrence of Rhinocerotina in the fossil record (one-horned rhinocerotine *Gaindatherium* cf. *browni*; Bugti Hills, Pakistan: 22.6 Ma; Antoine et al., 2013). (iii) The split between *Rhinoceros unicornis* and *Rhinoceros sondaicus* was calibrated as occurring between 1.9 Ma and 5.3 Ma. The minimum bound is supported by the earliest occurrence of both species in the fossil record (Antoine et al., 2021). A morphology-based phylogenetic analysis of Rhinocerotina, with a comprehensive sample within *Rhinoceros*, retrieves the following topology: (*R. sondaicus,*(*R. sinensis*,(*R. unicornis*,(*R. kendengindicus*, *R. platyrhinus*)))), with the first occurrence of *Rhinoceros platyrhinus* estimated at around the Pliocene–Pleistocene transition (2.58 Ma; Antoine et al., 2021). The maximum age constraint coincides with the earliest estimated age of *Rhinoceros* from the same morphology-based phylogenetic analyses. (iv) The split between *Ceratotherium simum* and *Diceros bicornis* was calibrated to occur between 5.3 Ma and 7.3 Ma. This range is consistent with records of *Diceros bicornis* recognized in upper Miocene deposits (> 5.3 Ma) at Lothagam (Kenya, 6.54-5.2 Ma; Brown and McDougall, 2011) and Albertine (Uganda, 7.25-5.3 Ma; Pickford et al., 1993).

Molecular data tend to identify the timing of the early radiation of mammals as being older than fossil evidence (e.g., O'Leary et al., 2013; dos Reis et al., 2014). For this reason, we performed analyses that included a fifth node-calibration on the age of the crown of all extant Perissodactyla. This calibration was a soft maximum bound at 66 Ma with a 0.01 prior probability of an older age. The age of this bound is based on the absence of unambiguous crown placental mammals before this time. All analyses were repeated after excluding this calibration for comparison (Figure S5A).

Additional molecular dating analyses were performed by including data from the tapir genome (Figure S5A), therefore including representatives of all living families of Perissodactyla (Equidae, Tapiridae, and Rhinocerotidae). The addition of these data was followed by filtering loci for biased rate estimates as described above, which led to a dataset of similar size (3,163 orthologs with 7,325,631 nucleotides). The species tree topology in these augmented analyses included the tapir as the sister to Rhinocerotidae, while the horse remained as the outgroup. We included a sixth calibration in addition to those described above, this time on the timing of the split between Rhinocerotidae and Tapiridae. The calibration had a hard minimum bound and soft maximum bound (0.025 prior probability on older ages), ranging between 54 Ma and 64 Ma. The minimum age is based on the earliest appearance of Tapiroidea in the fossil record shortly after the Paleocene-Eocene transition, around 54 Ma, with *Vastanolophus holbrooki* and *Cambaylophus vastanensis* in Asia (Rose et al., 2014) and 53.5 Ma with *Heptodon* in North America (Vandenberghe et al., 2012). The maximum age is based on the earliest record of Notoungulata at ~65 Ma (Tiupampa, Bolivia; Muizon and Ladevèze, 2020). The extinct South American orders Notoungulata and Litopterna form a clade, sister to Perissodactyla within Panperissodactyla (Buckley, 2015; Welker et al., 2015; Westbury et al., 2017), therefore suggesting a similar, or earlier, age for stem Perissodactyla.

Bayesian dating analyses were performed using a GTR+Γ substitution model and an uncorrelated -gamma relaxed clock model as implemented in MCMCtree, part of PAML v4.8 (Yang 2007). We further addressed heterogeneity in molecular evolutionary processes by partitioning the molecular clock and substitution models into each of the three codon positions (three partition subsets). We improved the efficiency of the analysis using approximate Bayesian computation (Thorne et al., 1998). The posterior distribution was estimated using Markov chain Monte Carlo (MCMC) samples. After a burn-in phase of 10^5 MCMC steps, samples were drawn

every 10^3 MCMC steps over a total of 10^7 steps. We verified convergence to the stationary distribution by comparing parameter estimates from four independent runs. Effective sample sizes were verified to be above 200 for all estimated parameters.

Alignment tool selection

The BWA-ALN algorithm (Li and Durbin, 2009) is commonly used to align such short reads generated in aDNA studies to reference genomes, owing to its stable performance for ultra-short length (typically shorter than 100 bp). However, simulation data showed read mapping success rate of the relatively longer reads (60 bp and 80 bp) dropped about three fold for high divergence reads (3% divergence level) using BWA-ALN (Parks and Lambert, 2015). We therefore compared the performance of BWA-ALN against BWA-MEM (Li, 2013), on the sequencing data obtained for the four historical/ancient samples (Figure S5B). For BWA-ALN, we disabled read seeding (-l 1024) to enhance error tolerance, while we marked shorter split hits as secondary alignments and removed them from the outputs for the BWA-MEM method.

Gene flow

To explore for potential gene flow between the different rhinoceros species, we computed D-statistics that assesses genetic affinities between taxa based on patterns of shared derived alleles (Green et al., 2010). For a given tree of topology (((H1, H2), H3), O), where O represents an outgroup, under the null hypothesis of no gene flow, shared derived alleles between H1 and H3, or H2 and H3, can only derive from incomplete lineage sorting (ILS) and are expected to be symmetric between the two pairs. Therefore, ancient gene flow is assumed between a pair in which imbalanced aggregation of shared derived alleles are detected. However, we should be aware of the limitations of D-statistics analysis between highly divergent lineages - certain assumptions could be violated in cases such as repeated or independent mutations (Prüfer et al., 2012) and differences in lineage specific mutation rates (Zheng and Janke, 2018). We independently mapped all raw reads from each rhinoceros species and the horse to three different reference genomes (greater one-horned rhinoceros, white rhinoceros and Sumatran rhinoceros) and then calculated D-statistics using the doAbbababa function module in ANGSD (Korneliussen et al., 2014) specifying the horse as the outgroup ancestral allele and the following filtering parameters: 1) use only sites where the reads of the outgroup all have the same base; 2) minimum base quality and mapping quality of 20; 3) use only the transversion sites, all individuals have a minimum depth of 3, and maximum depth of 70; 4) a block size of 5 Mb to estimate standard errors using jackknife procedure.

We further assessed whether less efficient mapping of the horse outgroup to the rhinoceros reference genomes may be driving some of our conflicting D-statistics results that differed based on the reference genome choice. We did this by calculating the pairwise distance of all rhinoceros species to the outgroup horse three times using all individuals mapped to the three different reference genomes. We calculated pairwise distances in 1 Mb sliding windows across all scaffolds > 1 Mb in size using a consensus base call approach (-doIBS 2) in ANGSD. We additionally applied the following filters; only include sites that have coverage in all individuals, a minimum base quality and mapping quality of 20, and only consider transversion differences.

Heterozygosity estimation

We estimated genome-wide heterozygosity for all the rhinoceros species based on a Site Frequency Spectrum (SFS), of which one diploid individual will generate three allelic states: homozygous ancestral allele state (AA), heterozygotic state (AB) and homozygous derived allele state (BB). Therefore, the whole genome level heterozygosity rate can be calculated as AB/(AA+AB+BB). We applied the doSaf function (-doSaf 1) module in ANGSD (version 0.924) (Korneliussen et al., 2014) to calculate genome-wide heterozygosity for each rhinoceros species with parameters of: unique mapping, a minimum mapping quality of 20, a minimum base quality of 20, a minimum and maximum depth value of 1/3 and 2 times of the average genome depths, respectively, for each species. For the ancient and historic samples (Javan, Merck's, woolly rhinoceros and Siberian unicorn), transition sites were removed (-noTrans 1) to eliminate the potential influences of DNA damage derived from cytosine deamination.

Demographic inference and Runs of Homozygosity (ROH) estimation

We applied the Pairwise Sequentially Markovian Coalescent (PSMC) method (Li and Durbin, 2011) to infer the history of population size changes in the ancestors of all the eight rhino species. For the species for which genomes were *de novo* assembled, we aligned ~30× genome coverage of the shotgun reads to recover the diploid genome heterozygosity information with depth filters of $\geq 1/3$ and $\leq 2 \times$ of the average depths. Then, PSMC was used to estimate the distribution of the time to the most recent common ancestor (TMRCA) between the two alleles across all chromosomes using the density information of heterozygous sites across each diploid genome. Changes in effective population size (N_e) were inferred assuming a substitution rate (μ) of 2.2×10^{-8} substitutions/site/generation and a generation time (g) of 25 years for the African rhinoceroses (Figure 3A; Moodley et al., 2020) and a μ of 2.34×10^{-8} and g of 12 years for the other rhinoceros species (Figures 3B and 3C; Dinerstein and McCracken, 1990; Lord et al., 2020; Mays et al., 2018). The consistency of the demographic results was tested by performing 100 bootstrap replicates as shown in Figure 4. ROH segments were recorded by summarizing genome regions of which the TMRCA dated to a recent time period (best fit K value of ≤ 2) following the method in Palkopoulou et al. (2015).

Genetic load and identification for rhinoceros specific frameshift mutation

First, gene annotation files of the *de novo* sequenced genomes (white, black, Sumatran and greater one-horned rhinoceroses) were imported into SnpEFF (Cingolani et al., 2012) to create a gene function database. Then, we obtained the heterozygous sites for each rhinoceros species using BCFTOOLS (Li, 2011) requiring: minimum and maximum depth of 1/3 and 2x of the average depth; mapping and base quality $\geq$ 20; variance quality $\geq$ 30; read supporting number of minor allele account for $\geq$ 20% of the total depth. After that, functional classes of the variances were estimated using SnpEFF (Cingolani et al., 2012) with default settings. We trimmed 10 bp from the 5′ and 3′ ends of each read for the Merck's rhinoceros and Siberian unicorn to alleviate the impact of DNA damages, because both the samples showed abnormal transition/transversion ratios - as high as 5.367 without end trimming (Figure S4B). However, as the transition/transversion ratio of the Merck's rhinoceros stayed abnormally high after trimming, which is consistent with previous reports of its DNA damage pattern (Figure S4 in Cappellini et al., 2019), we excluded it from further comparisons

To compare the rate of missense and loss-of-function mutations in rhinoceros to that of other mammals we obtained published re-sequencing data for 402 mammalian genomes from 30 species and mapped these to the phylogenetically closest available reference genome for each species (detailed in Table S4) using BWA-MEM v0.7.17 (Li, 2013). We then obtained and filtered variant calls for each individual using GATK HaplotypeCaller v3.8 following the "short variant discovery best-practices guidelines including "hard filtering" (Auwera et al., 2013). Additionally, we only retained within-species bi-allelic sites, and removed all indels and sites found at a frequency that was either less than one third, or greater than three times, the genome-wide autosomal coverage (Li, 2014).

To investigate the putative genetic background of unique rhinoceros biological adaptations, we explored frameshift mutations shared across the rhinoceros lineages. We inspected the frameshift mutations in the alignment results generated by the EXONERATE analysis mentioned above. First, all the frameshift mutations with their location information (Gene IDs and Exon IDs) were extracted out for rhinoceros species with *de novo* assemblies being available (black, white, Sumatran and greater one-horned rhinoceroses). Then, we filtered out the frameshift mutations that share the same locations and exist in all four rhinoceros species, and recorded their gene information for further examination.

Test of heterozygosity, PSMC and ROH estimation for the non-modern rhino samples

Since four of the rhinoceros genomes were generated from historic and ancient samples, data could only be generated from them by short read re-sequencing. Furthermore, given they represent historical or ancient samples, they are expected to contain damaged DNA (principally cytosine deamination (Pääbo, 1989)) manifested as elevated C > T/G > A transitions at the 5′ and 3′ ends of sequence reads (Briggs et al., 2007; Brotherton et al., 2007). Therefore, we tested the viability of restricting a number of the analyses performed to only the transversions, including heterozygosity estimation, historical demographical inference, and ROH estimation (Figures S3 and S4). Also, the influences of the other factors such as reference selection and sequencing depth were evaluated. For the ROH estimation, we applied the ROH inference method to the four rhinoceros species with *de novo* genome assemblies, using the datasets both with and without transitions. For the PSMC estimation, we mapped short reads representing various data volumes (different genome coverages, 11×, 17× and 35×) from the black rhinoceros onto the white rhinoceros genome to test whether it is possible to obtain accurate demographic history results using short read data from species that have a closely-related reference genome available.

QUANTIFICATION AND STATISTICAL ANALYSIS

Analyses related to genome assembly, assembly quality evaluation, short read alignment and evolutionary relationship inferences can be found in the Method details section.

ADDITIONAL RESOURCES

The results of some additional tests could be useful in understanding the sequence dataset for the historic/ancient rhinoceros samples and the comparative genome analyses, and can be found at: https://github.com/liushanlin/rhinoceros-comparative-genome/blob/main/additional_resources.md

Supplemental figures
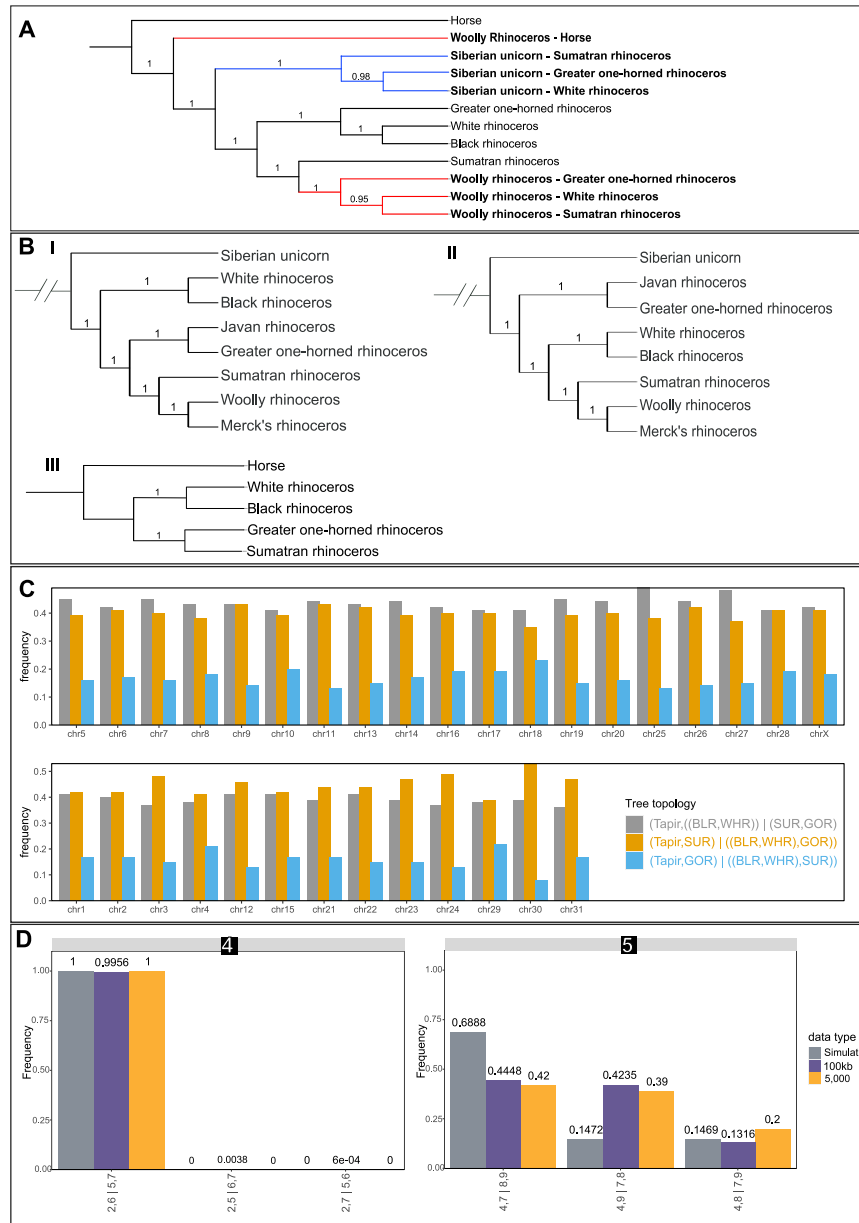
Figure S1. Phylogenetic relationship inferences, related to Figure 2

(A): Cladogram generated using different rhinoceros genomes as mapping references. Red branches represent different consensus sequences for the woolly rhinoceros generated when using different reference genomes for the mapping (second species named in each tip label), while blue represent the Siberian unicorn (*Elasmotherium sibiricum*). The cladogram shows a clear effect of reference bias when short reads are mapped on to a very distantly-related genome in order to obtain the consensus sequences. For example, while the position of the woolly rhinoceros is correct when mapped to either the white, Sumatran or greater one-horned rhinoceros, it is placed erroneously sister to all other rhinoceroses when mapped against the horse genome. Note that relationships were inferred under the multi-species coalescent as implemented in the software ASTRAL, which does not provide branch lengths in units of molecular substitutions nor terminal lengths, and thus we only show a tree topology. Branch labels indicate local posterior probabilities.

(B): Tree topologies obtained by mapping the Siberian unicorn against different reference genomes (I: white rhinoceros; II: greater one-horned rhinoceros) and using only the genomes that were *de novo* assembled (III). Branch labels indicate local posterior probabilities estimated in ASTRAL.

(C): The distribution of gene trees across the genome using the horse's chromosome ID as reference. The upper panel shows chromosomes that placed majority support on the species tree shown in Figure 2, while the bottom panel displays chromosomes with a greater frequency of an alternative gene tree.

(D): Comparison of the frequency of the three bipartitions between 100 kb sliding windows and 5,000 bp subregions. Branches 4 and 5 correspond to the tree in panel C in Figure 2.
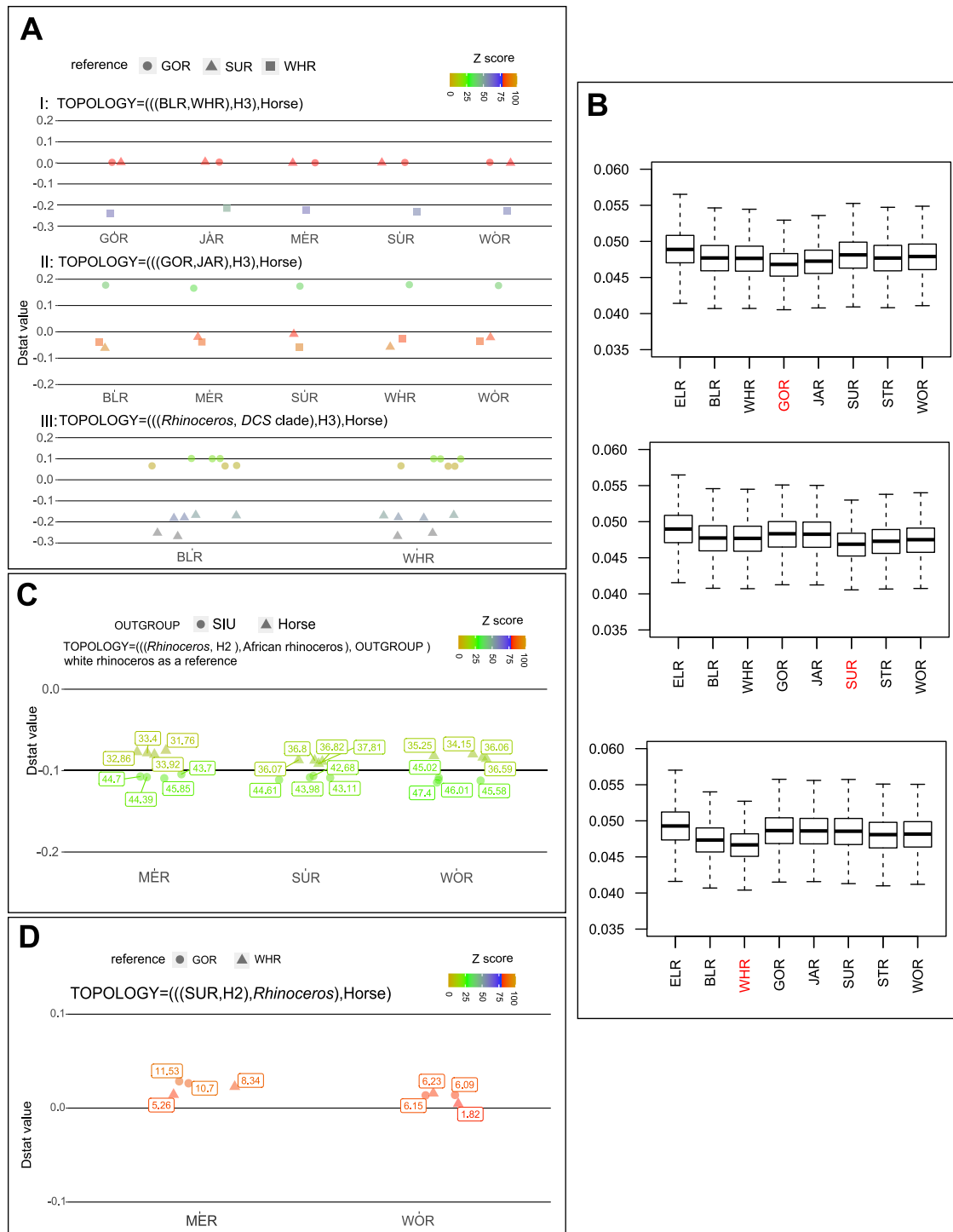
Figure S2. D-statistic analyses, related to Figure 2

GOR: greater one-horned rhinoceros; SUR: Sumatran rhinoceros; WHR: white rhinoceros; SIU: Siberian unicorn; BLR: black rhinoceros; JAR: Javan rhinoceros; MER: Merck's rhinoceros; WOR: woolly rhinoceros. Each dot represents a test using the topology and species in the subtitles. For example, (((GOR, SUR), BLR), Horse) represents one of the topologies in Panel A-III and is shown as one of the dots with the x axis of BLR. For a preset topology (((H1,H2),H3),H4), a negative D-statistics value means a closer relationship between H1 and H3 compared to that of H2 and H3, while a positive value means that H2 is closer to H3 than H1 is. Z score reflects the significance of the test with a value > 3 being considered as statistically significant.

(A): D-statistics score inference bias generated from reference selection. Species names on the x axis represent different H3 settings in the topology presets. It shows biases in the D-statistics scores that may have arisen as a consequence of the choice of reference genome against which the data was mapped prior to the

(legend continued on next page)

analysis. For example, when the species used as the mapping reference is placed at the H1 position, in every case, we observed an excess of shared derived alleles between the H2 and H3 species. We suggest that the deep divergence between H1 and H4 means that only more conserved regions, and regions more similar to the mapping reference will map. This will artificially make the alleles in the reference individual look more similar to the outgroup ancestral allele, resulting in more ABBA patterns and ultimately the biased D-statistics results that we see. (B) showed that regardless of the reference genome selected (i.e GOR, SUM, or WHR), a decreased pairwise distance between the horse and the reference genome (and the sister species of the reference to a lesser extent), relative to the other non-reference rhinoceros species. Taken together, using either the H1 or H2 individual as the mapping reference for this analysis will result in biased and likely incorrect inferences. Therefore, to avoid such obvious biases caused by the lower mapping efficiency of the outgroup horse genome, we only use D-statistics to infer gene flow when the reference genome is either in the H3 position, the sister species to the H3, or absent from the test entirely;

(B): The whole genome pairwise genetic distance between horse and the different rhinoceros species, when initially mapped against different rhinoceros reference genomes (indicated as red color, GOR (upper), SUR (middle) and WHR (bottom)). y axis represents the pairwise distance to horse;

(C): Excess of genetic affinities between Rhinoceros species (GOR and JAR) and the African rhinoceros (WHR and BLR). Species names on the x axis represent different H2 settings in the topology preset;

(D): Excess of genetic affinity between Rhinoceros and the two extinct DCS clade members (MER and WOR), but not with the Sumatran rhinoceros using two mapping references (GOR and WHR). Species names on the x axis represent different H2 settings in the topology preset.
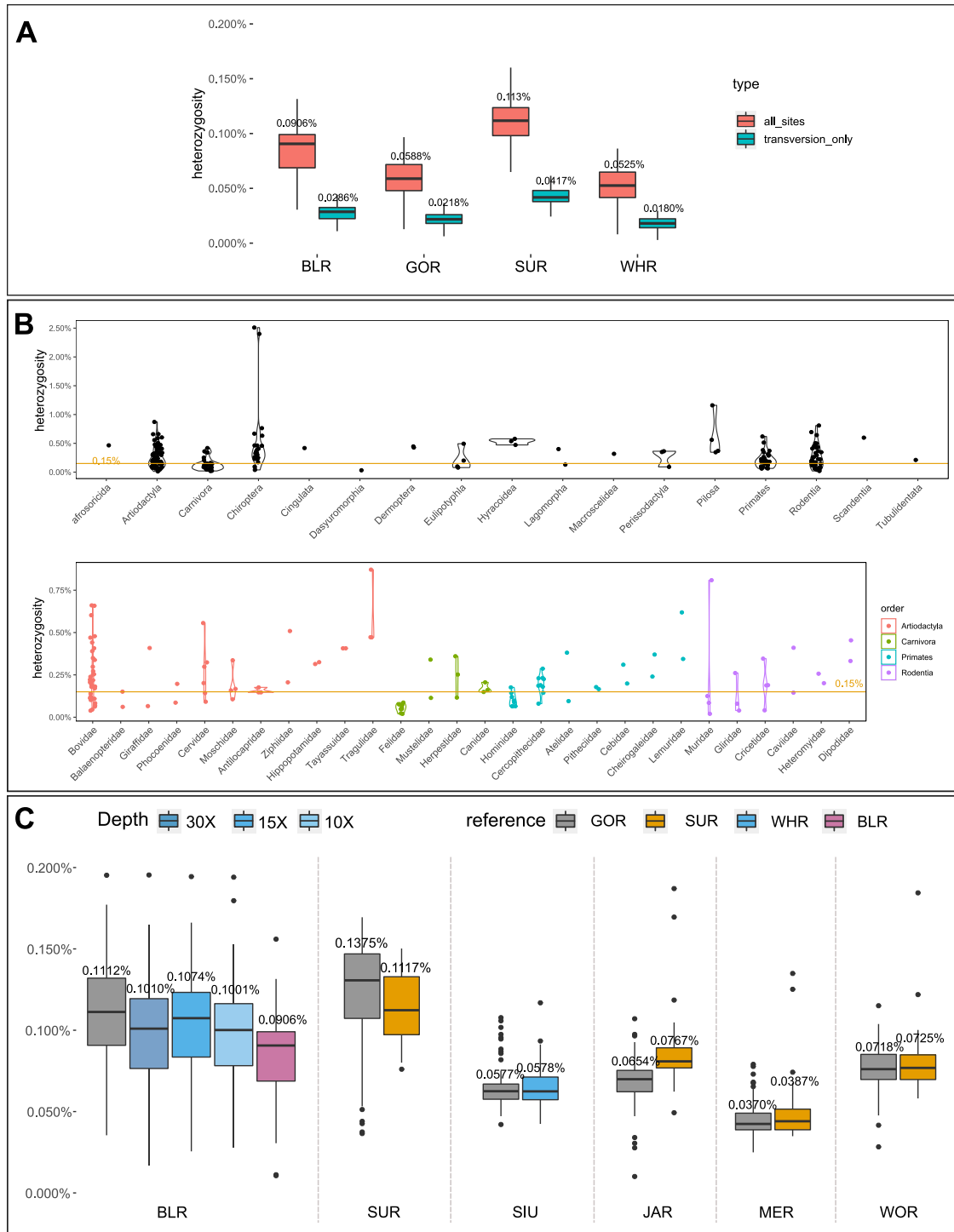
Figure S3. Heterozygosity estimations, related to Figure 3

GOR: greater one-horned rhinoceros; SUR: Sumatran rhinoceros; WHR: white rhinoceros; BLR: black rhinoceros; SIU: Siberian unicorn; JAR: Javan rhinoceros; MER: Merck's rhinoceros; WOR: woolly rhinoceros.

(A): Difference between heterozygosity ratio estimated taking into account all the sites and only the transversion sites. Numbers represent the median values for each group. For each box, the central line represents median value, and box limits represent first and third quartile, and whisker extends from the hinge to their largest values, but < 1.5*IQR (Interquartile range).

(B): Whole genome heterozygosity levels in different taxonomic groups. Note that all the humans and rhinoceros were excluded in the current figure. The yellow horizontal line represents a low heterozygosity of 0.15% similar to that of the rhinoceroses. In the bottom panel we only include families in Artiodactyla, Carnivora,
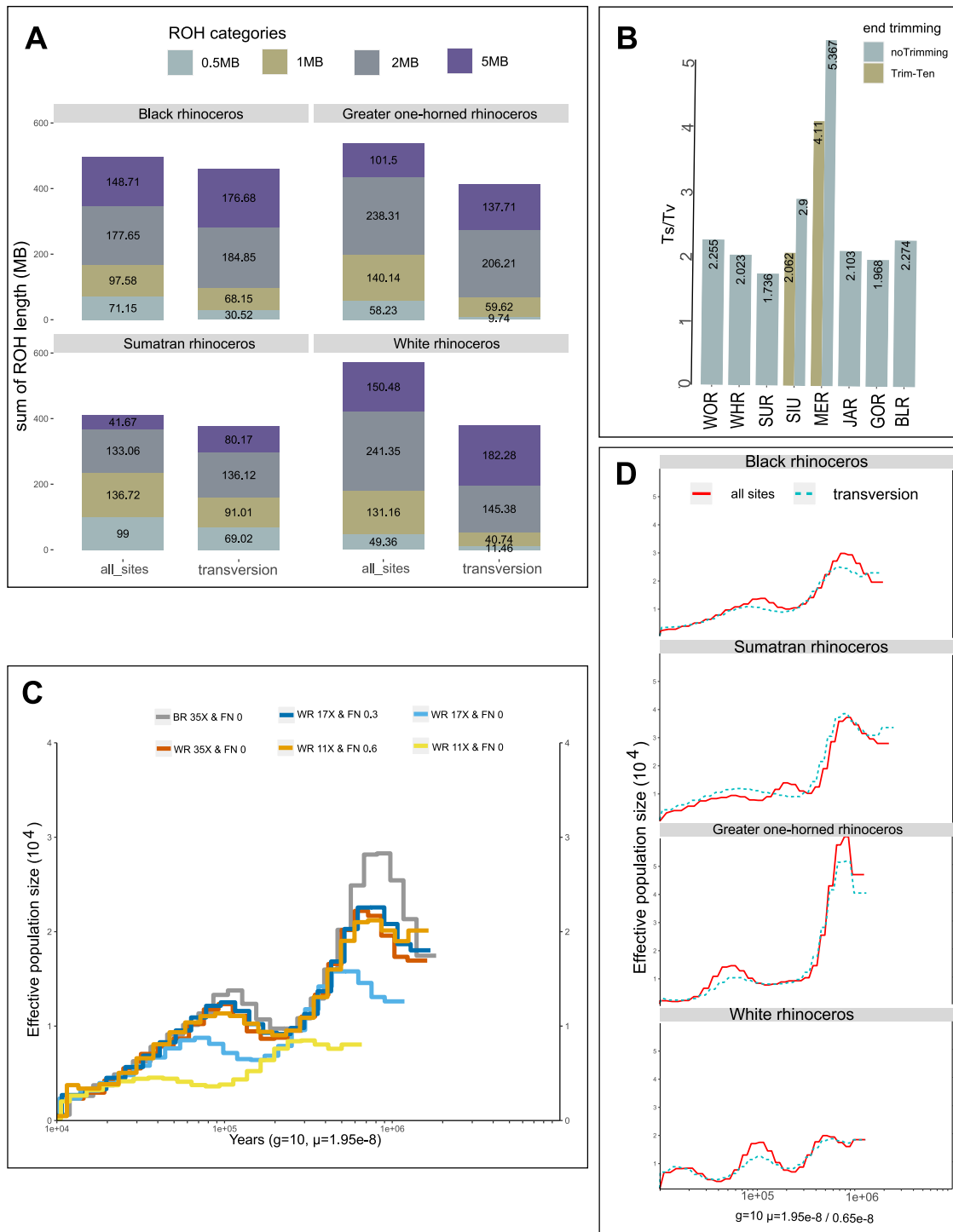
Primates and Rodentia that have multiple species of low heterozygosity level (upper panel), and families that have > 2 species in our dataset (Table S3 for all the details).

(C): Factors influencing heterozygosity ratio estimation. Numbers represent the median values for each group. Influence of genome coverage (Depth of 30×, 15× and 10× that indicated by different shades of blue) was tested by mapping reads from the black rhinoceros onto the reference genome of white rhinoceros. We noticed that different reference genomes contributed little to the variation of heterozygosity estimation, although when compared to the heterozygosity rate estimated using its own genome as a reference, using other closely related species tends to augment the estimation to a small degree and the inflation worsen when mapped to a higher divergent reference genome.

A: ROH categories — 0.5MB, 1MB, 2MB, 5MB

Black rhinoceros, Greater one-horned rhinoceros, Sumatran rhinoceros, White rhinoceros. sum of ROH length (MB); all_sites vs transversion.

B: end trimming — noTrimming, Trim-Ten. Ts/Tv.

C: BR 35X & FN 0, WR 35X & FN 0, WR 17X & FN 0.3, WR 11X & FN 0.6, WR 17X & FN 0, WR 11X & FN 0. Effective population size (10^4) vs Years (g=10, µ=1.95e-8).

D: Black rhinoceros, Sumatran rhinoceros, Greater one-horned rhinoceros, White rhinoceros — all sites, transversion. Effective population size (10^4). g=10 µ=1.95e-8 / 0.65e-8

Figure S4. Tests for ROH, PSMC, and gene effect estimations, related to Figures 3, 4 and 5

(A): ROH estimation comparison between the usage of all sites versus only transversions. This comparison shows that using transversions only leads to the detection of fewer ROH segments in total but tends to enlarge the ROH segments to longer stretches, and the differences are more obvious for the two rhinoceros species with relatively lower whole genome heterozygosity levels - the greater one-horned and white rhinoceroses;

(B): Transition/Transversion (Ts/Tv) ratio of different rhinoceros species. Trim-Ten (dark khaki) represent the Ts/Tv ratios calculated based on reads that were trimmed 10 bp from both the 5′ and 3′ end, while noTrimming (light blue) represent the rates obtained without read trimming;

(C): Demographic history estimated for black rhino. BR refers to black rhinoceros as reference genome, while WR refers to white rhinoceros as reference genome. FN means the false negative rate set for the PSMC curve plotting. The 35×, 17× and 11× represent the genome coverages for each test. It shows that adequate coverage (~30×) short reads, or inadequate coverage with proper FN settings can generate reliable demographic estimation;

(D): Demographic trajectories of different rhinoceros species. Results were estimated using both all heterozygous sites with a substitution rate of 1.95×10^{-8} per generation, and transversion heterozygous sites with a substitution rate of 0.65×10^{-8} per generation. We observed that the transversion only heterozygosity sites can also deliver comparable demographic results.
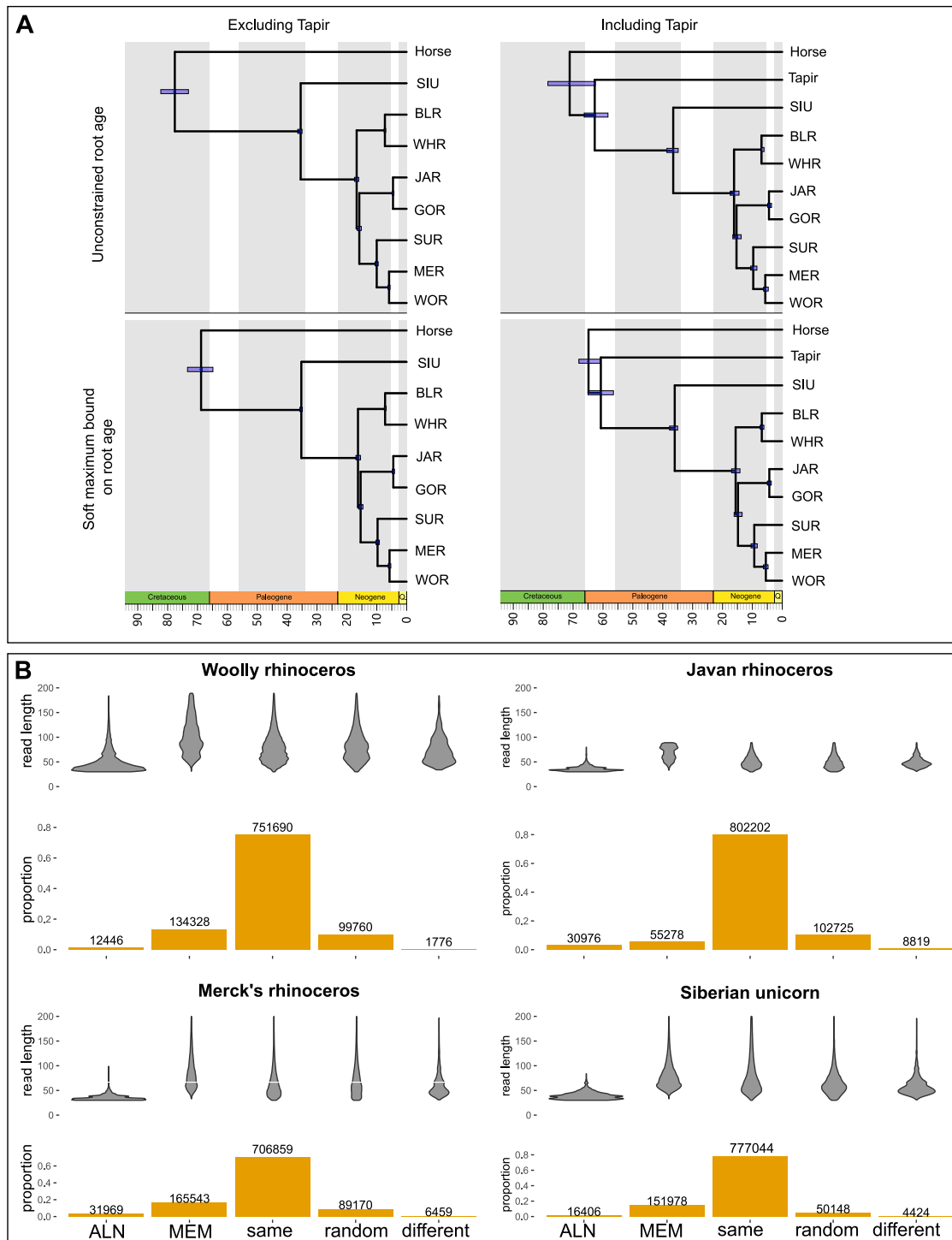
Figure S5. Analyses of molecular dating and alignment tool selection, related to Figure 2 and STAR Methods

(A): Molecular dating using different datasets and time-calibrations. The x-axes show the times of divergence in units of Million years ago (Mya). GOR: greater one-horned rhinoceros; SUR: Sumatran rhinoceros; WHR: white rhinoceros; SIU: Siberian unicorn; BLR: black rhinoceros; JAR: Javan rhinoceros; MER: Merck's rhinoceros; WOR: woolly rhinoceros.

(B): Comparison between the bwaALN and bwaMEM algorithms based on aligning one million randomly selected reads against their closely related reference genomes. The taxa used as reference for Siberian unicorn, Javan, woolly and Merck's rhinoceros are white, greater one-horn and Sumatran rhinoceroses, respectively. ALN and MEM on the x axis represent the number of reads that mapped successfully (Mapping Quality value ≥ 10) using the corresponding tools but failed with the other, and "random" represents the reads that had mapping quality values < 10 using both tools. "same" and "different" represent the number of reads that were mapped onto the same and different locations, respectively, using the two methods.