Article

Million-year-old DNA sheds light on the genomic history of mammoths

https://doi.org/10.1038/s41586-021-03224-9

Received: 3 July 2020

Accepted: 11 January 2021

Published online: 17 February 2021

Check for updates

Tom van der Valk[1,2,3,17]✉, Patrícia Pečnerová[2,4,5,17], David Díez-del-Molino[1,2,4,17], Anders Bergström[6], Jonas Oppenheimer[7], Stefanie Hartmann[8], Georgios Xenikoudakis[8], Jessica A. Thomas[8], Marianne Dehasque[1,2,4], Ekin Sağlıcan[9], Fatma Rabia Fidan[9], Ian Barnes[10], Shanlin Liu[11], Mehmet Somel[9], Peter D. Heintzman[12], Pavel Nikolskiy[13], Beth Shapiro[14,15], Pontus Skoglund[6], Michael Hofreiter[8], Adrian M. Lister[10], Anders Götherström[1,16,18] & Love Dalén[1,2,4,18]✉

Temporal genomic data hold great potential for studying evolutionary processes such as speciation. However, sampling across speciation events would, in many cases, require genomic time series that stretch well back into the Early Pleistocene subepoch. Although theoretical models suggest that DNA should survive on this timescale[1], the oldest genomic data recovered so far are from a horse specimen dated to 780–560 thousand years ago[2]. Here we report the recovery of genome-wide data from three mammoth specimens dating to the Early and Middle Pleistocene subepochs, two of which are more than one million years old. We find that two distinct mammoth lineages were present in eastern Siberia during the Early Pleistocene. One of these lineages gave rise to the woolly mammoth and the other represents a previously unrecognized lineage that was ancestral to the first mammoths to colonize North America. Our analyses reveal that the Columbian mammoth of North America traces its ancestry to a Middle Pleistocene hybridization between these two lineages, with roughly equal admixture proportions. Finally, we show that the majority of protein-coding changes associated with cold adaptation in woolly mammoths were already present one million years ago. These findings highlight the potential of deep-time palaeogenomics to expand our understanding of speciation and long-term adaptive evolution.

The recovery of genomic data from specimens that are many thousands of years old has improved our understanding of prehistoric population dynamics, ancient introgression events and the demography of extinct species[3–5]. However, some evolutionary processes occur over timescales that have often been considered beyond the temporal limits of ancient DNA research. For example, many present-day mammal and bird species originated during the Early and Middle Pleistocene[6,7]. Palaeogenomic investigations of their speciation process would thus require recovery of ancient DNA from specimens that are at least several hundreds of thousands of years old.

Mammoths (*Mammuthus* sp.) appeared in Africa approximately five million years ago (Ma), and subsequently colonized much of the Northern Hemisphere[8,9]. During the Pleistocene epoch (2.6 Ma to 11.7 thousand years ago (ka)), the mammoth lineage underwent evolutionary changes that produced the southern mammoth (*Mammuthus meridionalis)* and steppe mammoth (*Mammuthus trogontherii*), which later gave rise to the Columbian mammoth (*Mammuthus columbi*) and woolly mammoth (*Mammuthus primigenius*)[10]. Although the exact relationships among these taxa are uncertain, the prevailing view is that the Columbian mammoth evolved during an early colonization of North America about 1.5 Ma and that the woolly mammoth first appeared in northeastern Siberia about 0.7 Ma[8,10]. Mammoths similar to *M. trogontherii* (and considered conspecific with it) inhabited Eurasia from at least around 1.7 Ma; the last populations went extinct in Europe about 0.2 Ma[8].

To investigate the origin and evolution of woolly and Columbian mammoths, we recovered genomic data from three mammoth molars from northeastern Siberia that date to the Early and Middle Pleistocene (Fig. 1a, Extended Data Figs. 1, 2). These molars originate from the well-documented and fossiliferous Olyorian Suite of northeastern Siberia[11], which has been dated using rodent biostratigraphy tied to the global sequence of palaeomagnetic reversals as well as to correlated

[1]Centre for Palaeogenetics, Stockholm, Sweden. [2]Department of Bioinformatics and Genetics, Swedish Museum of Natural History, Stockholm, Sweden. [3]Department of Cell and Molecular Biology, National Bioinformatics Infrastructure Sweden, Science for Life Laboratory, Uppsala University, Uppsala, Sweden. [4]Department of Zoology, Stockholm University, Stockholm, Sweden. [5]Section for Computational and RNA Biology, Department of Biology, University of Copenhagen, Copenhagen, Denmark. [6]The Francis Crick Institute, London, UK. [7]Department of Biomolecular Engineering, University of California Santa Cruz, Santa Cruz, CA, USA. [8]Institute for Biochemistry and Biology, University of Potsdam, Potsdam, Germany. [9]Department of Biological Sciences, Middle East Technical University, Ankara, Turkey. [10]Department of Earth Sciences, Natural History Museum, London, UK. [11]College of Plant Protection, China Agricultural University, Beijing, China. [12]The Arctic University Museum of Norway, UiT – The Arctic University of Norway, Tromsø, Norway. [13]Geological Institute, Russian Academy of Sciences, Moscow, Russia. [14]Department of Ecology and Evolutionary Biology, University of California Santa Cruz, Santa Cruz, CA, USA. [15]Howard Hughes Medical Institute, University of California Santa Cruz, Santa Cruz, CA, USA. [16]Department of Archaeology and Classical Studies, Stockholm University, Stockholm, Sweden. [17]These authors contributed equally: Tom van der Valk, Patrícia Pečnerová, David Díez-del-Molino. [18]These authors jointly supervised this work: Anders Götherström, Love Dalén. ✉e-mail: tom.vandervalk@scilifelab.se; love.dalen@nrm.se

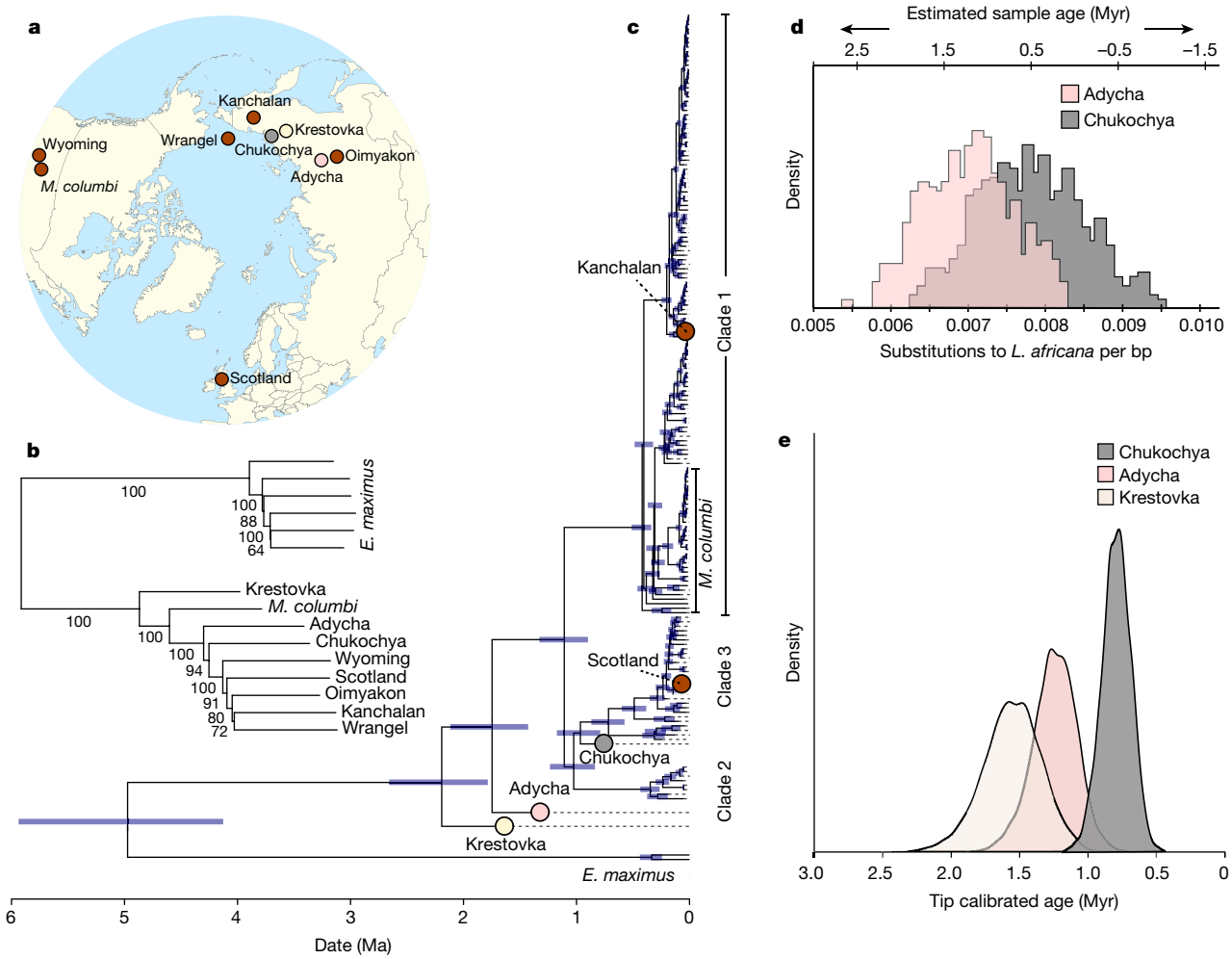

Fig. 1 | DNA-based phylogenies and specimen age estimates. a, Geographical origin of the mammoth genomes analysed in this study. **b**, Phylogenetic tree built in FASTME on the basis of pairwise genetic distances, assuming balanced minimum evolution using all nuclear sites as well as 100 resampling replicates (based on 100,000 sites each). **c**, Bayesian reconstruction of the mitochondrial tree, with the molecular clock calibrated using radiocarbon dates of ancient samples for which a finite radiocarbon date was available, as well as assuming a log-normal prior on the divergence between the African savannah elephant (not shown in the tree) and mammoths with a mean of 5.3 Ma. Blue bars reflect 95% highest posterior densities. Circles depict the position of the newly sequenced genomes. **d**, Densities for age estimates of the Adycha and Chukochya samples on the basis of autosomal divergence to African savannah elephant (*L. africana*). Owing to stochasticity among the tested blocks, a subset of genomic regions in the Krestovka and Adycha genomes are estimated as younger than the corresponding genomic region in the Wrangel mammoth genome, resulting in negative values. Myr, million years. **e**, Densities for age estimates of the Krestovka, Adycha and Chukochya samples on the basis of mitochondrial genomes, as inferred from the Bayesian mitochondrial reconstruction.

faunas with absolute dating from eastern Beringia (Extended Data Fig. 2, Supplementary Section 1). One of the specimens (which we refer to as 'Krestovka' on the basis of its find locality) is morphologically similar to the steppe mammoth (a species that was originally defined from the Middle Pleistocene of Europe (Supplementary Information section 1)), and was collected from Lower Olyorian deposits that have been dated to 1.2–1.1 Ma. The second specimen (referred to as 'Adycha'), which is also of *M. trogontherii*-like morphology (Supplementary Information section 1), is of a less-certain age within the Olyorian Suite (1.2–0.5 million years old). However, the morphology of the Adycha specimen (Extended Data Fig. 1) strongly suggests that it dates to the Early Olyorian, and probably to between 1.2 and 1.0 Ma. The third specimen (referred to as 'Chukochya') has a morphology consistent with being an early form of woolly mammoth (Extended Data Fig. 1) and was discovered in a section in which only Upper Olyorian deposits are exposed, which implies that it dates to 0.8–0.5 Ma (Supplementary Section 1).

We extracted DNA from the three molars using methods designed to recover highly degraded DNA fragments[12,13], converted the extracts into libraries[14] and sequenced these on Illumina platforms (Supplementary Information section 2, Supplementary Table 1). We merged the reads and mapped them against the African savannah elephant (*Loxodonta africana*) genome ('LoxAfr4')[15] and an Asian elephant (*Elephas maximus*) mitochondrial genome[16]. We found that the DNA recovered from the Early and Middle Pleistocene specimens was considerably more fragmented and had higher levels of cytosine deamination than DNA from permafrost-preserved samples dating to the Late Pleistocene subepoch (Extended Data Figs. 3, 4, Supplementary Information section 4). To circumvent this, we used conservative filters and an iterative approach that was designed to minimize spurious mappings of short reads (Supplementary Information section 5). This approach allowed us to recover complete (over 37× coverage) mitogenomes from all three specimens, and 49 million, 884 million and 3,671 million base pairs of nuclear genomic data for the Krestovka, Adycha and Chukochya specimens, respectively (Supplementary Table 3).

DNA-based age estimates

To estimate specimen ages using mitogenome data, we conducted a Bayesian molecular clock analysis that was calibrated using samples with finite radiocarbon dates (tip calibration) and a log-normal prior

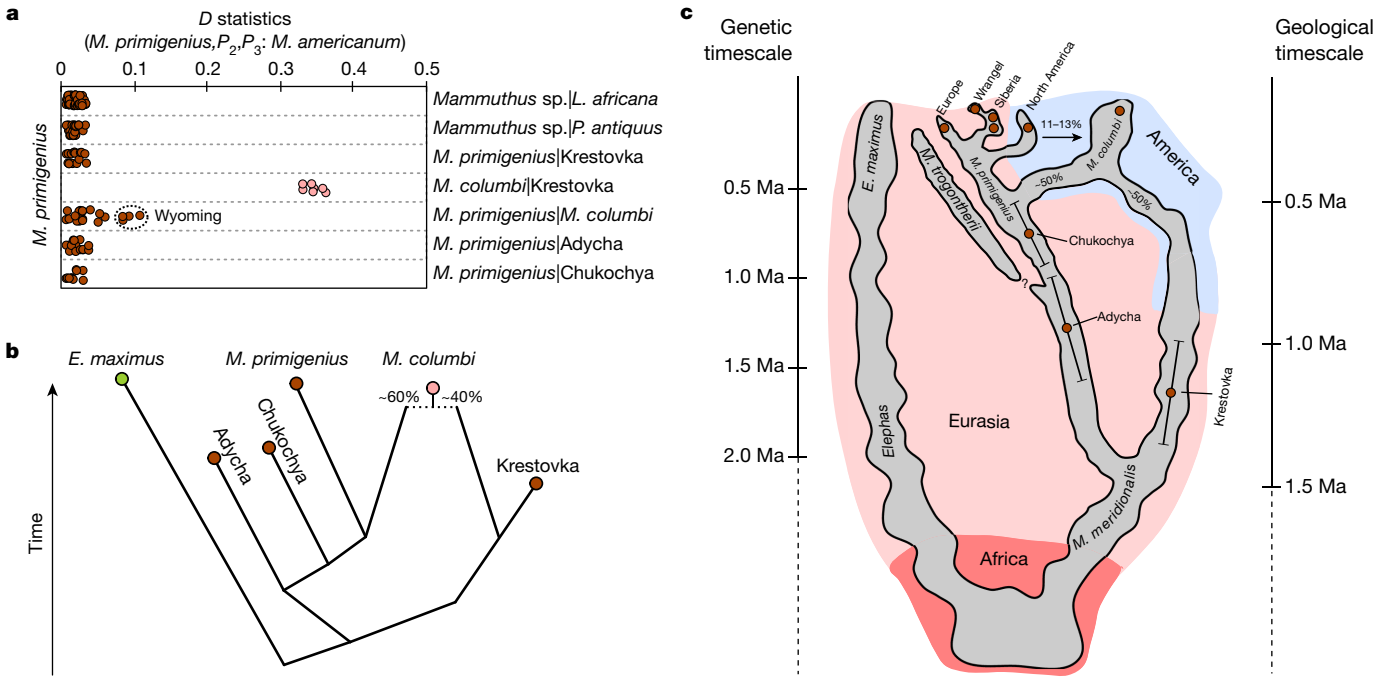

Fig. 2 | Inferred genomic history of mammoths. a, *D* statistics, in which each dot reflects a comparison involving one woolly mammoth genome and the two genomes depicted on the right, iterating through all possible sample combinations using the mastodon (*Mammut americanum*) as an outgroup. No elevated allele-sharing between any of the mammoth genomes and the reference (African savannah elephant) is observed, suggesting no pronounced reference biases in the Early and Middle Pleistocene genomes. A strong affinity between Columbian mammoths and the Krestovka sample is observed, as well as a relationship between the North American woolly mammoth (Wyoming) and the Columbian mammoth. The abbreviation *P. antiquus* denotes the straight-tusked elephant (*Palaeoloxodon antiquus*), and *Mammuthus* sp. refers to all mammoth specimens in this study. **b**, Best-fitting admixture graph model for one admixture event, suggesting a hybrid origin for the Columbian mammoth. **c**, Hypothesized evolutionary history of mammoths during the past 3 million years on the basis of currently available genomic data. Brown dots represent mammoth specimens for which genomic data have been analysed in this study; error bars represent 95% highest posterior density intervals from the mitogenome-based age estimates obtained for the three Early and Middle Pleistocene specimens. Arrows depict gene flow events identified from the autosomal genomic data. The European steppe mammoth (*M. trogontherii*) survived well into the later stages of the Middle Pleistocene, and we hypothesize that it most probably branched off from a common ancestor shared with the woolly mammoth about 1 Ma.

that assumed a genomic divergence between the African savannah elephant and mammoth lineages at 5.3 Ma[15] (root calibration). On the basis of this analysis, the specimens were estimated to date to 1.65 Ma (95% highest posterior density, 2.08–1.25 Ma), 1.34 Ma (1.69–1.06 Ma) and 0.87 Ma (1.07–0.68 Ma) for Krestovka, Adycha and Chukochya, respectively (Fig. 1c, e). We also used the autosomal genomic data to investigate the age of the higher-coverage Adycha (0.3×) and Chukochya (1.4×) specimens, by estimating the number of derived changes since their most recent common ancestor with the African savannah elephant (Supplementary Information section 6). We used an approach based on the accumulation of derived variants over time[17], assuming a constant mutation rate. This analysis suggested that the Adycha and Chukochya specimens date to 1.28 Ma (95% confidence interval, 1.64–0.92 Ma) and 0.62 Ma (95% confidence interval, 1.00–0.24 Ma), respectively (Fig. 1d). Although we caution that this analysis is based on low-coverage data and the confidence intervals are wide, these estimates are similar to those obtained from the mitochondrial data.

The DNA-based age estimates for the Chukochya and Adycha specimens are consistent with the geological age inferences that were independently derived from biostratigraphy and palaeomagnetism, whereas the molecular clock dating of the Krestovka specimen suggests an older age than that obtained from biostratigraphy. This could mean that the Krestovka specimen had been reworked from an older geological deposit or that the mitochondrial clock rate has been underestimated. However, the confidence intervals of the genetic and geological age estimates of the Krestovka specimen are separated by only 0.05 million years, and all estimates support an age greater than one million years.

A genetically divergent mammoth lineage

A phylogeny based on autosomal data shows that the three Early and Middle Pleistocene samples fall outside the diversity of all Eurasian mammoth genomes dating to the Late Pleistocene (Fig. 1b), including two woolly mammoth genomes from Europe (Scotland, dating to 48 ka) and Siberia (Kanchalan, dating to 24 ka) that were generated as part of this study. The phylogenetic positions of the Adycha and Chukochya specimens are consistent with their genomes being from a population directly ancestral to all Late Pleistocene woolly mammoths, whereas the Krestovka mammoth genome diverged before the split between the Columbian and woolly mammoth genomes (Fig. 1b). Similarly, Bayesian reconstruction of a mitogenome phylogeny that included 168 Late Pleistocene mammoth specimens[18,19] places the Early Pleistocene Krestovka and Adycha specimens as basal to all previously published mammoth mitogenomes, whereas the Middle Pleistocene Chukochya mitogenome is basal to one of the three clades that have previously been described for Late Pleistocene woolly mammoths[20] (Fig. 1c).

Estimates of sequence divergence times on the basis of both genome-wide and mitochondrial data indicate a deep split between the Krestovka specimen and all other mammoths analysed in this study. We estimate that the Krestovka mitogenome diverged from all other mammoth mitogenomes between 2.66 and 1.78 Ma (95% highest

posterior density) (Fig. 1c). We obtained a similar divergence time estimate (between 2.65 and 1.96 Ma based on the 95% confidence interval) from the autosomal data, but caution that this analysis is based on limited genomic data (Supplementary Information section 7). Moreover, estimates of relative divergence using $F(A|B)$ statistics[4] show that the Krestovka nuclear genome carries fewer derived alleles than any other mammoth genome at sites at which the high-coverage woolly mammoth genomes are heterozygous, which provides further support for the notion that the Krestovka mammoth lineage diverged after the split with Asian elephant but before any of the other mammoth genomes analysed here (Extended Data Fig. 5, Supplementary Information section 8).

Overall, these analyses suggest that two evolutionary lineages (that is, two isolated populations persisting through time) of mammoths inhabited eastern Siberia during the later stages of the Early Pleistocene. One of these lineages, which is represented by the Krestovka specimen, diverged from other mammoths before the first appearance of mammoths in North America. The second lineage comprises the Adycha specimen along with all Middle and Late Pleistocene woolly mammoths.

Origin of the Columbian mammoth

Several lines of evidence suggest that—compared to all other mammoths—the Columbian mammoth derives a much higher proportion of its ancestry from the lineage represented by the Krestovka mammoth. We performed analyses using D statistics[4], which revealed a strong signal of excess derived allele-sharing between the Columbian mammoth and the Krestovka specimen (Fig. 2a, Supplementary Information section 8). This is at odds with the average phylogenetic position of the Krestovka genome being basal to all other mammoth genomes, as under a scenario without subsequent admixture the D statistic would not deviate from zero. We further investigated this pattern using TreeMix[21]. Without modelling migration (admixture) events, none of the models fit the data (residuals > 10× s.e.). Instead, we observed a good fit when modelling one migration event (admixture weight = 42%, residuals < 2× s.e.) (Supplementary Information section 8), which indicates that part of the ancestry of the Columbian mammoth is derived from the Krestovka lineage.

To further assess the evolutionary context of the Krestovka lineage within the population history of mammoths, we used two complementary admixture graph model approaches[22,23]. We exhaustively tested all possible phylogenetic combinations relating the three ancient individuals with one Siberian woolly mammoth, one Columbian mammoth and one Asian elephant. We set the latter as outgroup, including only sites identified as polymorphic in six Asian elephant genomes to limit the effects of incorrectly called genotypes (Supplementary Information section 8). None of the graph models without admixture events provided a good fit to the data, thus ruling out a simple tree-like population history. By contrast, graph models with only one admixture event provided a perfect fit, explaining all 45 f_4-statistic combinations without significant outliers. On the basis of point estimates obtained from the two admixture graph model approaches, we estimate the Columbian mammoth to be the result of an admixture event in which 38–43% of its ancestry was derived from a lineage related to the Krestovka genome, and 57–62% from the woolly mammoth lineage (Fig. 2b, Extended Data Fig. 6).

We obtained additional support for the complex ancestry of the Columbian mammoth by using a hidden Markov model that aimed at identifying admixed genomic regions from an unknown source (that is, ghost admixture)[24] (Supplementary Information section 9). This analysis, which was done without including any of the Early and Middle Pleistocene specimens, suggested that roughly 41% of the Columbian mammoth genome originates from a lineage genetically differentiated from the woolly mammoth (Extended Data Fig. 7a). We subsequently built pairwise-distance phylogenetic trees for the genomic regions identified as being the result of ghost admixture and found them to be closely related to the Krestovka genome (Extended Data Fig. 7b, Supplementary Information section 9). By contrast, when excluding these regions, the remaining part of the Columbian mammoth genome falls within the diversity of Late Pleistocene woolly mammoths (Extended Data Fig. 7c, Supplementary Information section 9).

Finally, our D statistics analysis also identified higher levels of derived allele-sharing between the Columbian mammoth and a woolly mammoth from Wyoming (Fig. 2a). On the basis of f_4 ratios, we estimate 10.7–12.7% excess shared ancestry between these genomes (Supplementary Section 9), consistent with a previous study[15]. Because the Columbian mammoth carries a large proportion of Krestovka ancestry, gene flow from the Columbian mammoth into North American woolly mammoths would have resulted in a larger proportion of allele-sharing between Krestovka and the Wyoming woolly mammoth. Our finding of no excess allele-sharing between the Krestovka genome and any of the sequenced woolly mammoths—including the individual from Wyoming (Supplementary Table 7)—therefore indicates that this second phase of gene flow may have been unidirectional, from woolly mammoth into the Columbian mammoth. This implies that the composition of the genome of the Columbian mammoth (as identified in the D statistics, admixture graph models and ghost-admixture analysis) is the result of two admixture events, in which an initial approximately 50% contribution from each of the Krestovka and woolly mammoth lineages was followed by an additional approximately 12% gene flow from North American woolly mammoths (Fig. 2c).

Insights into mammoth adaptive evolution

The woolly mammoth evolved into a cold-tolerant, open-habitat specialist through a series of adaptive changes[8]. The antiquity of our genomes makes it possible to investigate when these adaptations evolved. To do this, we identified protein-coding changes for which all Late Pleistocene woolly mammoths carried the derived allele and all African savannah and Asian elephants carried the ancestral allele (n = 5,598) (Supplementary Table 8). Among the variants that could be called in the Early and Middle Pleistocene genomes, we find that 85.2% (782 out of 918) and 88.7% (2,578 out of 2,906) of the mammoth-specific protein-coding changes were already present in the genomes of Adycha (*M. trogontherii*-like) and Chukochya (early woolly mammoth), respectively (Supplementary Information section 10, Supplementary Table 9). Moreover, we did not detect significant differences in the ratio of shared nonsynonymous to synonymous sites among our sequenced Early, Middle and Late Pleistocene genomes (Supplementary Table 9). Thus, despite the transitions in climate and mammoth morphology at the onset of the Middle Pleistocene, we do not observe any marked change in the rate of protein-coding mutations during this time period.

Previous analyses have identified specific genetic changes that are thought to underlie a suite of woolly mammoth adaptations to the Arctic environment[25]. For these variants (n = 91), we assessed whether the Adycha and Chukochya genomes shared the same amino acid changes as those observed in Late Pleistocene woolly mammoths (Supplementary Table 10). We found that among genes that are possibly involved in hair growth, circadian rhythm, thermal sensation and white and brown fat deposits, the vast majority of coding changes were present in both the Adycha (87%) and Chukochya (89%) genomes (Supplementary Table 10). This suggests that Siberian *M. trogontherii*-like mammoths (that is, Adycha) had already developed a woolly fur as well as several physiological adaptations to a cold, high-latitude environment (Supplementary Information section 11). However, in one of the best-studied genes in the woolly mammoth (*TRPV3*, which encodes a temperature-sensitive transient receptor channel that is potentially involved in thermal sensation and hair growth[25]), we find that only two out of four amino acid changes identified in Late Pleistocene woolly

mammoths were present in the early woolly mammoth genome (Chukochya). This indicates that nonsynonymous changes in this gene occurred over several hundreds of thousands of years, rather than during a single brief burst of adaptive evolution.

Discussion

Our genomic analyses suggest that the Columbian mammoth is a product of admixture between woolly mammoths and a previously unrecognized ancient mammoth lineage, represented by the Krestovka specimen. Given the finding that each of these lineages initially contributed roughly half of their genome to this ancient admixture, we propose that the origin of the Columbian mammoth constitutes a hybrid speciation event[26]. This hybridization event appears not to have imparted any shift in the average molar morphology of North American populations[10], but can explain the mitochondrial–nuclear discordance in the Columbian mammoth[18], in which all known Columbian mammoth mitogenomes are nested within the mitogenome diversity of the woolly mammoth (Fig. 1c). On the basis of the mitogenome phylogeny, we estimate that the most recent common female ancestor of all Late Pleistocene Columbian mammoths lived approximately 420 ka (95% highest posterior density, 511–338 ka), providing a likely minimum date for when this hybridization event occurred (Fig. 1c). Because mammoths had already appeared in North America by 1.5 Ma, these findings imply that before the hybridization event North American mammoths belonged to the Krestovka lineage. Given the morphology of the Krestovka specimen, this corroborates a previously proposed model[10] that the earliest North American mammoths were derived from an *M. trogontherii*-like Eurasian ancestor, rather than originating from an expansion of the southern mammoth (*M. meridionalis*) into North America[27].

Our findings demonstrate that genomic data can be recovered from Early Pleistocene specimens, which opens up the possibility of studying adaptive evolution across speciation events. The mammoth genomes presented here offer a glimpse of this potential. Even though the transition from an *M. trogontherii*-like (Adycha) to woolly (Chukochya) mammoth represents a marked change in molar morphology (Extended Data Fig. 1), we do not observe an increased rate of genome-wide selection during this time period. Moreover, many key adaptations identified in Late Pleistocene mammoth genomes were already present in the Early Pleistocene Adycha genome. We thus find no evidence for an increased rate of adaptive evolution associated with the origin of the woolly mammoth. This is consistent with previous work that suggested that the major shift in habitat and morphology of mammoths happened earlier, between *M. meridionalis*-like and *M. trogontherii*-like mammoths[8,10].

The retrieval of DNA that is more than one million years old confirms previous theoretical predictions[1] that the ancient genetic record can be extended beyond what has been previously shown. We anticipate that the additional recovery and analysis of Early and Middle Pleistocene genomes will further improve our understanding of the complex nature of evolutionary change and speciation. Our results highlight the value of perennially frozen environments for extending the temporal limits of DNA recovery, and hint at a future deep-time chapter of ancient DNA research in which specimens from high latitudes will have an important role.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-021-03224-9.

1. Allentoft, M. E. et al. The half-life of DNA in bone: measuring decay kinetics in 158 dated fossils. *Proc. R. Soc. Lond. B* **279**, 4724–4733 (2012).
2. Orlando, L. et al. Recalibrating *Equus* evolution using the genome sequence of an early Middle Pleistocene horse. *Nature* **499**, 74–78 (2013).
3. Skoglund, P. et al. Origins and genetic legacy of Neolithic farmers and hunter-gatherers in Europe. *Science* **336**, 466–469 (2012).
4. Green, R. E. et al. A draft sequence of the Neandertal genome. *Science* **328**, 710–722 (2010).
5. Palkopoulou, E. et al. Complete genomes reveal signatures of demographic and genetic declines in the woolly mammoth. *Curr. Biol.* **25**, 1395–1400 (2015).
6. Weir, J. T. & Schluter, D. Ice sheets promote speciation in boreal birds. *Proc. R. Soc. Lond. B* **271**, 1881–1887 (2004).
7. Lister, A. M. The impact of Quaternary Ice Ages on mammalian evolution. *Phil. Trans. R. Soc. Lond. B* **359**, 221–241 (2004).
8. Lister, A. M., Sher, A. V., van Essen, H. & Wei, G. The pattern and process of mammoth evolution in Eurasia. *Quat. Int.* **126–128**, 49–64 (2005).
9. Werdelin, L. & Sanders, W. J. (eds) *Cenozoic Mammals of Africa* (Univ. California Press, 2010).
10. Lister, A. M. & Sher, A. V. Evolution and dispersal of mammoths across the Northern Hemisphere. *Science* **350**, 805–809 (2015).
11. Repenning, C. A. Allophaiomys *and the Age of the Olyor Suite, Krestovka Sections, Yakutia* (US Government Printing Office, 1992).
12. Dabney, J. et al. Complete mitochondrial genome sequence of a Middle Pleistocene cave bear reconstructed from ultrashort DNA fragments. *Proc. Natl Acad. Sci. USA* **110**, 15758–15763 (2013).
13. Briggs, A. W. et al. Removal of deaminated cytosines and detection of in vivo methylation in ancient DNA. *Nucleic Acids Res.* **38**, e87 (2010).
14. Meyer, M. & Kircher, M. Illumina sequencing library preparation for highly multiplexed target capture and sequencing. *Cold Spring Harb. Protoc.* **2010**, db.prot5448 (2010).
15. Palkopoulou, E. et al. A comprehensive genomic history of extinct and living elephants. *Proc. Natl Acad. Sci. USA* **115**, E2566–E2574 (2018).
16. Rohland, N. et al. Proboscidean mitogenomics: chronology and mode of elephant evolution using mastodon as outgroup. *PLoS Biol.* **5**, e207 (2007).
17. Meyer, M. et al. A high-coverage genome sequence from an archaic Denisovan individual. *Science* **338**, 222–226 (2012).
18. Chang, D. et al. The evolutionary and phylogeographic history of woolly mammoths: a comprehensive mitogenomic analysis. *Sci. Rep.* **7**, 44585 (2017).
19. Pečnerová, P. et al. Mitogenome evolution in the last surviving woolly mammoth population reveals neutral and functional consequences of small population size. *Evol. Lett.* **1**, 292–303 (2017).
20. Barnes, I. et al. Genetic structure and extinction of the woolly mammoth, *Mammuthus primigenius*. *Curr. Biol.* **17**, 1072–1075 (2007).
21. Pickrell, J. K. & Pritchard, J. K. Inference of population splits and mixtures from genome-wide allele frequency data. *PLoS Genet.* **8**, e1002967 (2012).
22. Patterson, N. et al. Ancient admixture in human history. *Genetics* **192**, 1065–1093 (2012).
23. Leppälä, K., Nielsen, S. V. & Mailund, T. admixturegraph: an R package for admixture graph manipulation and fitting. *Bioinformatics* **33**, 1738–1740 (2017).
24. Skov, L. et al. Detecting archaic introgression using an unadmixed outgroup. *PLoS Genet.* **14**, e1007641 (2018).
25. Lynch, V. J. et al. Elephantid genomes reveal the molecular bases of woolly mammoth adaptations to the Arctic. *Cell Rep.* **12**, 217–228 (2015).
26. Mallet, J. Hybrid speciation. *Nature* **446**, 279–283 (2007).
27. Lucas, S. G., Morgan, G. S., Love, D. W. & Connell, S. D. The first North American mammoths: taxonomy and chronology of early Irvingtonian (Early Pleistocene) *Mammuthus* from New Mexico. *Quat. Int.* **443**, 2–13 (2017).

Methods

No statistical methods were used to predetermine sample size. The experiments were not randomized, and investigators were not blinded to allocation during experiments and outcome assessment.

Morphometry of mammoth molars

Mammoth molars were measured according to a previously described method[10] (Supplementary Information section 1). Samples considered are as follows: *M. meridionalis*, about 2.0 Ma, Upper Valdarno, Italy (type locality) ($n = 34$); *M. trogontherii*, about 0.6 Ma, Süssenborn, Germany (type locality) ($n = 48$); and *M. primigenius*, Late Pleistocene of north-eastern Siberia (Russia) and Alaska (USA) ($n = 28$). Early ($n = 8$) and Late ($n = 15$) Olyorian samples are from localities in the Yana–Kolyma lowland (Lower Olyorian Suite is about 1.2–0.8 Ma; Upper Olyorian Suite is 0.8–0.5 Ma) (Extended Data Fig. 2). Early to early Middle Pleistocene samples (about 1.5–0.5 Ma) from North America are from Old Crow (Yukon, Canada), Leisey Shell Pit 1A and Punta Gorda (both in Florida, USA), and the Ocotillo Formation (California, USA) (combined, $n = 16$). Original data have previously been published[10], along with further details on sites and collections.

DNA extraction and sequencing

Samples from Early and Middle Pleistocene mammoth molars (Krestovka, Adycha and Chukochya specimens) as well as Late Pleistocene samples (Scotland and Kanchalan specimens) were processed in dedicated ancient DNA laboratories following standard ancient DNA practices (Supplementary Information section 2). Following DNA extraction[12], we constructed double- or single-stranded Illumina libraries[14,28], which were treated to remove uracil caused by post-mortem cytosine deamination[13]. We subsequently sequenced these libraries using Illumina platforms, generating from 200 to 2,350 million paired-end reads (2×50 or 2×150 bp) per specimen (Supplementary Table 1).

Sequence data processing and mapping

We combined our sequence data with previously published genomic data from elephantids[15] (Supplementary Table 2). For the five samples sequenced in this study, we trimmed adapters and merged paired-end reads using SeqPrep v.1.1[29], initially retaining reads either ≥25 bp (Krestovka, Adycha and Chukochya specimens) or ≥30 bp (Scotland and Kanchalan specimens), and with a minor modification in the source code that enabled us to choose the best base-quality score in the merged region instead of aggregating the scores[5] (Supplementary Information section 3). For genomic data from the straight-tusked elephant and the Scotland and Kanchalan mammoths (which had been treated with Afu uracil-DNA glycosylase (UDG), leaving post-mortem DNA damage at the ends of the molecules (Supplementary Tables 2, 3)), we removed the first and last two base pairs from all reads before mapping. The merged reads were mapped to a composite reference, consisting of the African savannah elephant nuclear genome (LoxAfr4), woolly mammoth mitogenome (DQ188829) and the human genome (hg19) using BWA aln v.0.7.8 with deactivated seeding (-l 16,500), allowing for more substitutions (-n 0.01) and up to two gaps (-o 2)[30,31]. The human genome was included as a decoy to filter out spurious mappings in genomic conserved regions[32]. Next, we removed PCR duplicates from the alignments using a custom Python script[5]. After obtaining initial quality metrics for the genomes, we removed reads <35 base pairs from the BAM files using samtools v.1.10[33] and awk for all remaining analysis (Supplementary Section 4).

Ancient DNA authenticity and quality assessment

All ancient genomes were treated to reduce post-mortem DNA damage. For the most ancient samples (Krestovka, Adycha and Chukochya), we took several steps to assess the authenticity and quality of the data (Supplementary Information section 4). First, only reads that mapped uniquely to nonrepetitive regions of the LoxAfr4 reference and had a mapping quality ≥30 were retained; reads that mapped equally well to the human genome reference (hg19) in our composite reference were removed to reduce possible biases caused by contaminant human reads[32]. Second, we used a method based on the rate of mismatches per base pair to the reference to assess the rate of spurious mappings for all reads between 20 and 35 bp and at 5-bp intervals between 35 and 50 bp (Supplementary Information section 4). This enabled us to identify a sample-specific minimum read length cut-off, above which we consider reads to be correctly mapped and endogenous (Supplementary Information section 4, Supplementary Table 3). On the basis of this, we applied the longest sample-specific cut-off (≥35 bp, for the Krestovka specimen) for all samples. We used mapDamage v.2.0.6[34] to obtain read-length distributions for all ancient samples. Finally, an assessment of cytosine deamination profiles at CpG sites, which are unaffected by UDG treatment[13], was done using the platypus option in PMDtools (https://github.com/pontussk/PMDtools)[35]. A full set of ancient DNA quality statistics are available in Supplementary Tables 1–3.

Allele sampling

To minimize coverage-related biases, all subsequent analyses were based on pseudo-haploidized sequences that were generated by randomly selecting a single high-quality base call at each autosomal genomic site using ANGSD v.0.921[36]. For base-calling, we considered only reads ≥ 35 bp, a mapping and base quality ≥ 30 and reads without multiple best hits (-uniqueOnly 1). Finally, we masked all sites within repetitive regions as identified with RepeatMasker v.4.0.7[37], CpG sites, sites with more than two alleles among all individuals and sites with coverage above the 95th percentile of the genome-wide average, to reduce false calls from duplicated genomic regions.

Reconstruction of mitogenomes, tip-dating and mitochondrial DNA phylogeny

Mitochondrial genomes for the five newly sequenced samples were assembled using MIA[38] with the Asian elephant (NC_005129)[16] mitogenome as reference for Adycha, Krestovka and Chukochya specimens, and the mammoth mitogenome (NC_007596) as reference for the Late Pleistocene woolly mammoth samples from Scotland and Kanchalan, restricting the input reads to those ≥ 35 bp for each (Supplementary Section 5). This yielded mitochondrial assemblies with coverage of 37.8×, 47.5× and 77.1× for the Adycha, Krestovka and Chukochya specimens, and 99.6× and 179.5× for the Scotland and Kanchalan samples, respectively. These assemblies were then aligned using Muscle v.3.8.31[39] together with previously published elephantid mitogenomes[18,19,40]. Following alignment partitioning, the HKY model with a gamma-distributed rate heterogeneity[41] and a proportion of invariant sites or just a proportion of invariant sites, was identified as best-fitting for each alignment partition using jModelTest v.2.1.10[42] (Supplementary Information section 5). To estimate the age of the three oldest *Mammuthus* samples (Adycha, Krestovka and Chukochya), we performed a Bayesian reconstruction of the phylogenetic tree using BEAST v.1.10.4[43]. We calibrated the molecular clock using tip ages for all ancient samples with a finite radiocarbon date, as well as a log-normal prior of 5.3 Ma on the genetic divergence of *Loxodonta* and *Elephas–Mammuthus* as obtained from previous genomic studies[15] (Supplementary Table 4). In addition, we tested for an older divergence (7.6 Ma) between *Loxodonta* and *Mammuthus* that is more consistent with the fossil record[16] (Supplementary Information section 5). For both priors, we used a standard deviation of 500,000 years. We assumed a strict molecular clock and the flexible skygrid coalescent model[44] to account for the complex cross-generic demographic history of the included taxa. The ages of all samples beyond the limit of radiocarbon dating were estimated by sampling from log-normal distributions with priors based on stratigraphic context and previous genetic studies, using two Markov chain Monte Carlo (MCMC) chains of 100 million generations, sampling

every 10,000 and discarding the first 10% as burn-in (Supplementary Table 5, Supplementary Information section 5).

Genetic dating on the basis of autosomal data

Age estimates for the Adycha and Chukochya specimens (the Krestovka specimen was excluded as too few autosomal bases were available for this analysis) were estimated on the basis of autosomal data following a previously described method[17], using the American mastodon (*Mammut americanum*; an outgroup to all elephantids) and the African savannah and Asian elephant genomes as outgroups. We inferred the ancestral state for a given base in the African savannah elephant reference genome by requiring that the alignments of the mastodon, two African savannah elephants and five Asian elephants are present and identical at that nucleotide. We used the high-coverage and radiocarbon-dated Wrangel Island woolly mammoth genome as a calibration point[5]. Each difference to the ancestral state was then counted for the Wrangel genome and the focal *Mammuthus* genome for all sites at which both genomes had a called base. We calculated the relative age of each individual as $(n_W - n_M)/n_W$, on the basis of the number of derived changes in the Wrangel genome (n_W) and the other *Mammuthus* genome (n_M), using an assumed divergence time of 5.3 million years[15] to the common ancestor of African savannah elephant and woolly mammoth. Age variance estimates were calculated in windows of 5 Mb and we computed bootstrap confidence intervals as 1.96× s.e. around the date estimates (Supplementary Information section 6).

Nuclear genetic relationships and phylogeny

We reconstructed phylogenetic trees on the basis of the whole-genome identical-by-state matrix for all individuals using the doIBS function in ANGSD. We calculated pairwise genetic distances between individuals using the full dataset, as well as 100 resampling replicates based on 100,000 sites each. Second, we obtained the phylogenetic tree using a balanced minimum evolution method as implemented in FASTME[45] (Fig. 1b, Supplementary Information section 7). Next, we inferred relative population split times using an approach that examines single-nucleotide polymorphic positions that are heterozygous in an individual from one population and measures the fraction of these sites at which a randomly sampled allele from an individual of a second population carries the derived variant, polarized by an outgroup ($F(A|B)$ statistics)[4]. We ascertained heterozygous sites in three high-coverage genomes—*E. maximus* and *M. primigenius* Oimyakon and Wrangel[5]—using the SAMtools v.1.10[33] mpileup command and bcftools. We only included single-nucleotide polymorphisms with a quality ≥ 30, and filtered out all single-nucleotide polymorphisms in repetitive regions, within 5 bp of insertions and/or deletions, at CpG sites and sites below 1/3 or above 2× the genome-wide average coverage. For each of the *Mammuthus* genomes, we then estimated the proportion of sites for which a randomly drawn allele at the ascertained heterozygous sites matches the derived state.

D statistics, f_4 statistics, AdmixtureGraphs and TreeMix

We first used Admixtools v.5[22] to calculate *D* statistics and f_4 statistics for all possible quadruple combinations of samples iterating through the three different groups (P_1, P_2 and P_3) on the basis of randomly sampled alleles, conditioning on all sites that are polymorphic among the six Asian elephant genomes[22]. The mastodon was used as an outgroup in all comparisons (Supplementary Tables 6, 7). Direct estimates of genomic ancestries using f_4 ratios were additionally calculated for specific pairs in AdmixTools[22] (Supplementary Information section 9). Second, we used the admixturegraph R package[23] to assess the genetic relationship among the *Mammuthus* genomes using admixture graph models, fitting graphs to all possible f_4 statistics involving a given set of genomes. To resolve the relationships of the Adycha, Krestovka and Chukochya individuals within the population history of mammoths, we exhaustively tested all 135,285 possible admixture graphs (with up to 2 admixture events) relating these 3 individuals, 1 woolly mammoth (Wrangel), 1 Columbian mammoth and 1 Asian elephant, setting the latter as outgroup (Supplementary Information section 8). We repeated the admixturegraph analysis using the above-described f_4 statistic with qpBrute[46], which in addition enabled us to estimate shared genetic drift and branch lengths using f_2 and f_3 statistics. At each step, insertion of a new node was tested at all branches of the graph, except the outgroup branch. In cases in which a node could not be inserted without producing f_4 outliers (that is, $|Z| \geq 3$), all possible admixture combinations were also attempted. The resulting list of all fitted graphs was then passed to the MCMC algorithm implemented in the admixturegraph R package, to compute the marginal likelihood of the models and their Bayes factors. Finally, we estimated genetic relationships and admixture among the *Mammuthus* samples using TreeMix v.1.12[21]. We first estimated the allele frequencies among the randomly sampled alleles and subsequently ran the TreeMix model accounting for linkage disequilibrium by grouping sites in blocks of 1,000 single-nucleotide polymorphisms (-k 1,000) setting the *E. maximus* samples as root. Standard errors (-SE) and bootstrap replicates (-bootstrap) were used to evaluate the confidence in the inferred tree topology. After constructing a maximum-likelihood tree, migration events were added (−*m*) and iterated 10 times for each value of *m* (1–10) to check for convergence in the likelihood of the model as well as the explained variance following each addition of a migration event. The inferred maximum-likelihood trees were visualized with the in-built TreeMix R script plotting functions.

Introgression in the Columbian mammoth

We further tested for admixture in the Columbian and Scotland mammoths using a hidden Markov model[24]. This method identifies genomic regions within a given individual that possibly came from an admixture event with a distant lineage not present in the dataset, on the basis of on the distribution of private sites. In brief, we estimated the number of callable sites, the single-nucleotide polymorphism density (as a proxy for per-window mutation rate) and the number of private variants with respect to all other elephant genomes except Krestovka in 1-kb windows. We applied settings without gene flow, or with one gene flow event with starting probabilities and decoding described in Supplementary Information section 9. We tested for ghost admixture in the Columbian mammoth using sites private to the Columbian mammoth with respect to all other genomes in this study except Krestovka. We subsequently obtained fasta alignments for those autosomal regions identified as 'unadmixed' and 'ghost-admixed' in the Columbian mammoths by calling a random base at each covered position using ANGSD. Minimal evolution phylogenies were then obtained for both alignments as described in 'Nuclear genetic relationships and phylogeny'.

Genetic adaptations of the woolly mammoth

To investigate the timing of genetic adaptations in the woolly mammoth lineage, we used last v.1170[47] to build a chain file to lift over our sampled allele dataset mapped to LoxAfr4 to the annotated LoxAfr3 reference genome. Following construction of a reference index using lastdb *(*-P0 -uNEAR -R01), we aligned the two references using lastal (-m50 -E0.05 -C2). The alignment was converted to MAF format (last-split -m1) and finally to a chain file with the maf-convert tool (http://last.cbrc.jp/). The Picard Liftover tool (https://broadinstitute.github.io/picard/) was then used to lift over the identified variants to the LoxAfr3 reference. Using the African savannah elephant genome annotation (LoxAfr3.gff), we identified all amino acid changes in which all Late Pleistocene woolly mammoth genomes carry the derived state and all other elephantid genomes carry the ancestral allele using VariantEffectPredictor[48]. For all identified amino acid changes, we assessed the state (derived or ancestral) among the three oldest samples (Krestovka, Adycha and Chukochya) and the Columbian mammoth (Supplementary Tables 8–10). In addition, we conducted a Gene Ontology enrichment on all genes

for which the woolly mammoth genomes (including Chukochya and Adycha) are derived, using GOrilla[49]. Finally, we used PAML v.1.3.1[50] to identify genes that have potentially been under positive selection in Late Pleistocene woolly mammoths (Supplementary Table 11, Supplementary Information section 10).

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

All sequence data (in .fastq format) for samples sequenced in this study are available through the European Nucleotide Archive under accession number PRJEB42269. Previously published data used in this study are available under accession numbers PRJEB24361 and PRJEB7929.

Code availability

The custom code used in this study to evaluate read length cut-offs is available from GitHub (https://github.com/stefaniehartmann/readLengthCutoff).

28. Gansauge, M.-T. & Meyer, M. Single-stranded DNA library preparation for the sequencing of ancient or damaged DNA. *Nat. Protocols* **8**, 737–748 (2013).
29. John, J. S. SeqPrep: tool for stripping adaptors and/or merging paired reads with overlap into single reads. GitHub https://github.com/jstjohn/SeqPrep (2011).
30. Schubert, M. et al. Improving ancient DNA read mapping against modern reference genomes. *BMC Genomics* **13**, 178 (2012).
31. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. Preprint at https://arxiv.org/abs/1303.3997 (2013).
32. Feuerborn, T. R. et al. Competitive mapping allows for the identification and exclusion of human DNA contamination in ancient faunal genomic datasets. *BMC Genomics* **21**, 844 (2020).
33. Li, H. et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
34. Jónsson, H., Ginolhac, A., Schubert, M., Johnson, P. L. F. & Orlando, L. mapDamage2.0: fast approximate Bayesian estimates of ancient DNA damage parameters. *Bioinformatics* **29**, 1682–1684 (2013).
35. Skoglund, P. et al. Separating endogenous ancient DNA from modern day contamination in a Siberian Neandertal. *Proc. Natl Acad. Sci. USA* **111**, 2229–2234 (2014).
36. Korneliussen, T. S., Albrechtsen, A. & Nielsen, R. ANGSD: analysis of next generation sequencing data. *BMC Bioinformatics* **15**, 356 (2014).
37. Smit, A. F. A., Hubley, R. & Green, P. RepeatMasker Open-4.0, 2013–2015. http://www.repeatmasker.org (2015).
38. Green, R. E. et al. A complete Neandertal mitochondrial genome sequence determined by high-throughput sequencing. *Cell* **134**, 416–426 (2008).
39. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792–1797 (2004).
40. Meyer, M. et al. Palaeogenomes of Eurasian straight-tusked elephants challenge the current view of elephant evolution. *eLife* **6**, e25413 (2017).
41. Yang, Z. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *J. Mol. Evol.* **39**, 306–314 (1994).
42. Darriba, D., Taboada, G. L., Doallo, R. & Posada, D. jModelTest 2: more models, new heuristics and parallel computing. *Nat. Methods* **9**, 772 (2012).
43. Suchard, M. A. et al. Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. *Virus Evol.* **4**, vey016 (2018).
44. Gill, M. S. et al. Improving Bayesian population dynamics inference: a coalescent-based model for multiple loci. *Mol. Biol. Evol.* **30**, 713–724 (2013).
45. Lefort, V., Desper, R. & Gascuel, O. FastME 2.0: a comprehensive, accurate, and fast distance-based phylogeny inference program. *Mol. Biol. Evol.* **32**, 2798–2800 (2015).
46. Liu, L. et al. Genomic analysis on pygmy hog reveals extensive interbreeding during wild boar expansion. *Nat. Commun.* **10**, 1992 (2019).
47. Frith, M. C., Hamada, M. & Horton, P. Parameters for accurate genome alignment. *BMC Bioinformatics* **11**, 80 (2010).
48. McLaren, W. et al. The Ensembl variant effect predictor. *Genome Biol.* **17**, 122 (2016).
49. Eden, E., Navon, R., Steinfeld, I., Lipson, D. & Yakhini, Z. GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. *BMC Bioinformatics* **10**, 48 (2009).
50. Yang, Z. PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24**, 1586–1591 (2007).

Acknowledgements T.v.d.V., P.P., D.D.-d.-M., M.D. and L.D. acknowledge support from the Swedish Research Council (2012-3869 and 2017-04647), FORMAS (2018-01640) and the Tryggers Foundation (CTS 17:109). A.G. is supported by the Knut and Alice Wallenberg Foundation (1,000 Ancient Genomes project). A.B. and P.S. were supported by the Francis Crick Institute (FC001595), which receives its core funding from Cancer Research UK, the UK Medical Research Council and the Wellcome Trust. P.S. was supported by the European Research Council (grant no. 852558), the Wellcome Trust (217223/Z/19/Z) and the Vallee Foundation. M.H., J.A.T., I.B., A.M.L. and G.X. were supported by NERC (grant no. NE/J010480/1) and the ERC StG grant GeneFlow (no. 310763). B.S. and J.O. were supported by the US National Science Foundation (DEB-1754451). P.N. was supported by RFBR (grant no. 13-05-01128). The authors also acknowledge support from Science for Life Laboratory, the Knut and Alice Wallenberg Foundation, the National Genomics Infrastructure funded by the Swedish Research Council, and Uppsala Multidisciplinary Center for Advanced Computational Science for assistance with massively parallel sequencing and access to the UPPMAX computational infrastructure. N. Clark at the Hunterian Museum provided access to the Scotland mammoth sample. Finally, we thank our late friend and colleague A. Sher, who defined and described the Olyorian sequence, collected large quantities of fossil vertebrate material (including all of the Early and Middle Pleistocene specimens studied here) and consistently promoted multidisciplinary studies on his finds.

Author contributions L.D., A.M.L., B.S., M.H. and I.B. conceived the project. L.D., A.G., P.P. and D.D.-d.-M. designed the study together with P.N. and A.M.L. Laboratory work on Early and Middle Pleistocene samples was done by P.P., L.D., A.G. and M.D., and G.X. and J.A.T. conducted laboratory work on Late Pleistocene samples. P.P., T.v.d.V. and D.D.-d.-M. processed and mapped sequence data. T.v.d.V., S.H. and P.D.H. performed tests on DNA authenticity. T.v.d.V., J.O. and S.L. conducted phylogenetic and Treemix analyses. J.O. and T.v.d.V. computed genomic age estimates. T.v.d.V., A.B. and D.D.-d.-M. performed analyses on *D* statistics and f_4 statistics and admixture graph models. T.v.d.V. performed analyses on population structure, and ghost admixture. T.v.d.V., E.S., F.R.F. and M.S. performed analysis on selection. L.D., P.D.H., M.H., B.S., A.G., M.S., P.S., P.N. and A.M.L. provided advice on the bioinformatic analyses and/or helped to interpret the results. P.N. and A.M.L. provided morphological analyses as well as palaeontological and geological information. The manuscript was written by T.v.d.V., P.P., D.D.-d.-M., P.N. and L.D., with contributions from all co-authors.

Competing interests The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41586-021-03224-9.

Correspondence and requests for materials should be addressed to T.v.d.V. or L.D.

Peer review information *Nature* thanks Gloria Cuenca-Bescós, David Lambert, Krishna Veeramah and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.