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

Tom van der Valk[1,2,3*], Patrícia Pečnerová[2,4,5*], David Díez-del-Molino[1,2,4*], 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#], Love Dalén[1,2,4#]

1. Centre for Palaeogenetics, Svante Arrhenius väg 20C, SE-106 91 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, SE-106 91 Stockholm, Sweden
5. Section for Computational and RNA Biology, Department of Biology, University of Copenhagen, DK-2200 Copenhagen, Denmark
6. The Francis Crick Institute, London NW1 1AT, UK
7. Department of Biomolecular Engineering, University of California Santa Cruz, Santa Cruz, CA, USA
8. Institute for Biochemistry and Biology, University of Potsdam, 14476 Potsdam, Germany
9. Department of Biological Sciences, Middle East Technical University, Ankara, Turkey
10. Department of Earth Sciences, Natural History Museum, London SW7 5BD, UK.
11. College of Plant Protection, China Agricultural University, Beijing 100193, China
12. The Arctic University Museum of Norway, UiT - The Arctic University of Norway, 9037 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 96054 USA
16. Department of Archaeology and Classical Studies, Stockholm University, SE-106 91 Stockholm, Sweden

*) These authors contributed equally: Tom van der Valk, Patrícia Pečnerová, David Díez-del-Molino
#) These authors jointly supervised this work: Anders Götherström and Love Dalén
Correspondence: tom.vandervalk@scilifelab.se, love.dalen@nrm.se

Abstract

Temporal genomic data hold great potential for studying evolutionary processes, including speciation. However, sampling across speciation events would in many cases require genomic time series that stretch well into the Early Pleistocene (>1 million years). Although theoretical models suggest that DNA should survive on this timescale[1], the oldest genomic data recovered so far is from a 560-780 ka old horse specimen[2]. Here we report the recovery of genome-wide data from three Early and Middle Pleistocene mammoth specimens, 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 gave rise to the woolly mammoth, whereas the other represents a previously unrecognised lineage that was ancestral to the first mammoths to colonise North America. Our analyses reveal that the North American Columbian mammoth traces its ancestry to a Middle Pleistocene hybridisation 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 present already a million years ago. These findings highlight the potential of deep time palaeogenomics to expand our understanding of speciation and long-term adaptive evolution.

Main

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 time scales 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 (ka) old.

Mammoths (*Mammuthus* sp.) appeared in Africa approximately 5 million years ago (Ma) and subsequently colonised much of the Northern Hemisphere[8,9]. During the Pleistocene (2.6 Ma - 11.7 ka), the mammoth lineage underwent evolutionary changes that resulted in early species known as the southern (*Mammuthus meridionalis)* and steppe (*M. trogontherii*) mammoths, which later gave rise to the Columbian (*M. columbi*) and woolly (*M. primigenius*) mammoths[10]. Although the exact relationships among these taxa are uncertain, the prevailing view is that the Columbian mammoth evolved during an early colonisation of North America c. 1.5 Ma, whereas the woolly mammoth first appeared in northeastern Siberia c. 0.7 Ma[8,10]. *M. trogontherii*-like mammoths, considered to be a single species, inhabited Eurasia since at least c. 1.7 Ma, with the last populations going extinct in Europe at c. 0.2 Ma[8].

To investigate the origin and evolution of woolly and Columbian mammoths, we recovered genomic data from three northeastern Siberian mammoth molars dated to the Early and Middle Pleistocene (Fig. 1a; Extended Data Fig. 1; Extended Data Fig. 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 faunas with absolute dating from eastern Beringia (Extended Data Fig. 2, Supplementary Section 1). One of the specimens (Krestovka) is morphologically similar to the steppe mammoth, a species originally defined from the European Middle Pleistocene (Supplementary Section 1), and was collected from Lower Olyorian deposits that have been dated to 1.2 - 1.1 Ma. The second specimen (Adycha), which is also of *trogontherii*-like morphology (Supplementary Section 1), is of less certain age within the Olyorian (1.2 - 0.5 Ma). However, the morphology of the Adycha specimen (Extended data Fig. 1) strongly suggests that it dates to the Early Olyorian, 1.2 - 1.0 Ma. The third specimen (Chukochya) has a morphology consistent with an early form of woolly mammoth (Extended data Fig. 1) and was discovered in a section where only Upper Olyorian deposits are exposed, implying an approximate age of 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 Section 2; Supplementary Table 1). The reads were merged and mapped 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 Late Pleistocene permafrost samples (Extended Data Figs. 3, 4, Supplementary Section 4). To circumvent this, we used conservative filters and an iterative approach designed to minimise spurious mappings of short reads (Supplementary Section 5). This approach allowed us to recover complete (>37X coverage) mitogenomes from all three specimens, and 49, 884, and 3,671 million base pairs of nuclear genomic data for Krestovka, Adycha, and Chukochya, respectively (Supplementary Table 3).

DNA-based age estimates

To estimate specimen ages using mitogenome data, we conducted a Bayesian molecular clock analysis, calibrated using samples with finite radiocarbon dates (tip calibration) and a log-normal prior assuming a 5.3 Ma genomic divergence between the African elephant and mammoth lineages[15] (root calibration). This provided specimen age estimates of 1.65 Ma (95% HPD: 2.08-1.25 Ma), 1.34 (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 (0.3X) and Chukochya (1.4X) specimens by estimating the number of derived changes since their common ancestor with the African elephant (Supplementary Section 6). We used an approach based on the accumulation of derived variants over time[17], assuming a constant mutation rate. This resulted in inferred ages of 1.28 Ma (95% CI 1.64-0.92 Ma) for the Adycha specimen and 0.62 Ma (95% CI 1.00-0.24 Ma) for the Chukochya specimen (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 independently derived geological age inferences from biostratigraphy and palaeomagnetism, whereas molecular clock dating of the Krestovka specimen suggests an older age compared to 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 Ma, 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/Middle Pleistocene samples fall outside the diversity of all Late Pleistocene Eurasian mammoth genomes (Fig. 1b), including two woolly mammoth genomes from Europe (Scotland; 48 ka) and Siberia (Kanchalan; 24 ka) generated as part of this study. The phylogenetic positions of Adycha and Chukochya are consistent with these genomes being from a population directly ancestral to all Late Pleistocene woolly mammoths, whereas the Krestovka mammoth genome diverged prior to the split between 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 previously described for Late Pleistocene woolly mammoths[20] (Fig. 1c).

Estimates of sequence divergence times based on both genome-wide and mitochondrial data indicate a deep split between Krestovka 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% HPD, Fig. 1c). We obtained a similar divergence time estimate (95% CI 2.65 - 1.96 Ma) from the autosomal data, but caution that this analysis is based on limited genomic data (Supplementary 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 where the high-coverage woolly mammoth genomes are heterozygous, further supporting that it diverged after the split with Asian elephant but before any of the other mammoth genomes analysed here (Extended Data Fig. 5, Supplementary Section 8).

Overall, these analyses suggest that two evolutionary lineages (*i.e.* two isolated populations persisting through time) of mammoths inhabited eastern Siberia during the latter stages of the Early Pleistocene. One of these lineages, which is represented by the Krestovka specimen, diverged from other mammoths prior to 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

Intriguingly, 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. Analyses using D-statistics[4] revealed a strong signal of excess derived allele sharing between the Columbian mammoth and Krestovka (Fig. 2a, Supplementary Section 8). This is at odds with the average phylogenetic position of Krestovka being basal to all other mammoth genomes, since 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 >10x SE). Instead, we observed a good fit when modelling one migration event (admixture weight = 42%; residuals <2x SE) (Supplementary section 8), indicating that part of the Columbian mammoth's ancestry 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, only including sites identified as polymorphic in six Asian elephant genomes to limit the effects of incorrectly called genotypes (Supplementary Section 8). None of the graph models without admixture events provided good fits to the data, thus ruling out a simple tree-like population history. In contrast, graph models with just one admixture event provided a perfect fit, explaining all 45 f_4-statistic combinations without significant outliers. Based on the point estimates obtained from the two different admixture graph model approaches, the Columbian mammoth is estimated to be the result of an admixture event where 38-43% of its ancestry was derived from a lineage related to Krestovka, 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 employing a hidden Markov model aimed at identifying admixed genomic regions from an unknown source (*i.e.* ghost admixture)[24] (Supplementary 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 closely related to the Krestovka genome (Extended Data Fig. 7b, Supplementary Section 9). In 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 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). Based on f_4-ratios, we estimate 10.7-12.7% excess shared ancestry between these genomes (Supplementary Section 9), consistent with an earlier study[15]. Since 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 Columbian mammoth's genome, as identified in the D-statistics, admixture graph models, and ghost-admixture analysis, is the result of two admixture events, where an initial ~50% contribution from each of the Krestovka and woolly mammoth lineages was followed by an additional ~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 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 (*trogontherii*-like) and Chukochya (early woolly mammoth), respectively (Supplementary Section 10, Supplementary Table 9). Moreover, we did not detect significant differences in the ratio of shared non-synonymous versus 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 find that among genes 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 *trogontherii*-like mammoths (*i.e.* Adycha) had already developed a woolly fur as well as several physiological adaptations to a cold high-latitude environment (Supplementary Section 11). However, in one of the best studied genes in the woolly mammoth, *TRPV3*, which encodes a temperature-sensitive transient receptor channel, 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 non-synonymous 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 unrecognised 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 hybridisation event appears not to have imparted any shift in average molar morphology of North American populations[10], but can explain the mitochondrial-nuclear discordance in the Columbian mammoth[18] where all known Columbian mammoth mitogenomes are nested within the woolly mammoth's mitogenome diversity (Fig. 1c). Based on the mitogenome phylogeny, we estimate that the most recent common female ancestor of all Late Pleistocene Columbian mammoths lived approximately 420 ka (95% HPD 511 - 338 ka), providing a likely minimum age for when this hybridization event occurred (Fig. 1c). Since mammoths had already appeared in North America by 1.5 Ma, these findings imply that prior to the hybridisation event, North American mammoths belonged to the Krestovka lineage. Given the morphology of the Krestovka specimen, this corroborates the model proposed by Lister & Sher[10] that the earliest North American mammoths were derived from a *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, opening 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 *trogontherii*-like (Adycha) to woolly (Chukochya) mammoths represents a significant 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 suggesting that the major shift in habitat and morphology of mammoths happened earlier, between *meridionalis*-like and *trogontherii*-like mammoths[8,10].

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

References (Main)

1. Allentoft, M. E. *et al.* The half-life of DNA in bone: measuring decay kinetics in 158 dated fossils. *Proc. Biol. Sci.* **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. Biol. Sci.* **271**, 1881–1887 (2004).
7. Lister, A. M. The impact of Quaternary Ice Ages on mammalian evolution. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **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. *Quaternary International* vols 126-128 49–64 (2005).
9. *Cenozoic Mammals of Africa*. (University of 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*. (U.S. 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. U. S. A.* **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. U. S. A.* **115**, E2566–E2574 (2018).

16. Rohland, N. *et al.* Proboscidean mitogenomics: chronology and mode of elephant evolution using mastodon as outgroup. *PLoS Biol.* **5**, (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).

Figure legends

Fig. 1. DNA-based phylogenies and specimen age estimates. a, Geographic origin of the mammoth genomes analysed in this study. **b,** Phylogenetic tree built in FASTME based on 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 lognormal 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 samples Adycha and Chukochya based on autosomal divergence to African savannah elephant (*L. africana*) and **e,** Densities for age estimates of samples Krestovka, Adycha and Chukochya based on mitochondrial genomes as inferred from the Bayesian mitochondrial reconstruction.

Fig. 2. Inferred genomic history of mammoths. a, D-statistics where each dot reflects a comparison involving one woolly mammoth genome and one genome depicted on the right side of the panel (where *L. africana* = African savannah elephant, *P. antiquus* = straight-tusked elephant, *Mammuthus sp*. = all mammoth specimens in this study, *M. columbi* = Columbian mammoth, and *M. primigenius* = woolly mammoth), 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/Middle Pleistocene genomes. A strong affinity between Columbian mammoths and sample Krestovka is observed, as well as a relationship between the North American woolly mammoth (Wyoming) and the Columbian mammoth. **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 last 3 Ma, based on currently available genomic data. Brown dots represent mammoth specimens for which genomic data has been analysed in this study, with error bars representing 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 likely branched off from a common ancestor shared with the woolly mammoth at ~1 Ma.

Methods

Morphometry of mammoth molars

Mammoth molars were measured according to the method described in Lister & Sher[10] (Supplementary Section 1). Samples considered are as follows: *Mammuthus meridionalis*, ca. 2.0 Ma, Upper Valdarno, Italy (type locality) (n=34); *M. trogontherii*, ca. 0.6 Ma, Süssenborn, Germany (type locality) (n=48); *M. primigenius*, Late Pleistocene of North-East Siberia (Russia) and Alaska (USA) (n=28). Early (n=8) and Late (n=15) Olyorian samples are from localities in the Yana-Kolyma lowland (Early Olyorian is ~1.2 – 0.8 Ma, Late Olyorian is 0.8 – 0.5 Ma; Extended Data Fig. 2). North American Early to early Middle Pleistocene samples (ca. 1.5 – 0.5 Ma) are from Old Crow (Yukon, Canada), Leisey Shell Pit 1A and Punta Gorda (Florida, USA), and the Ocotillo Formation (California, USA) (combined n=16). Original data are from Lister & Sher[10], where further details on sites and collections can be found.

DNA extraction and sequencing

Samples from Early-Middle Pleistocene mammoth molars (Krestovka, Adycha, Chukochya) as well as Late Pleistocene samples (Scotland, Kanchalan) were processed in dedicated ancient DNA laboratories following standard ancient DNA practices (Supplementary Section 2). Following DNA extraction[12], we constructed double- or single-stranded Illumina libraries[14,28], which were treated to remove uracils 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 (2x 50 or 2x150 bp) per specimen (Supplementary Table 1).

Sequence data processing and mapping

We combined our sequence data with previously published genomic data from elephantids generated by Palkopoulou *et al*.[15] (Supplementary Table 2). For the five samples sequenced in this study, we trimmed adapters and merged paired-end reads using SeqPrep v1.1[29], initially retaining reads either ≥25 bp (Krestovka, Adycha, Chukochya) or ≥30 bp (Scotland, Kanchalan), and with a minor modification in the source code that allowed us to choose the best base quality score in the merged region instead of aggregating the scores[5] (Supplementary Section 3). For genomic data from the straight-tusked elephant, and the Scotland and Kanchalan mammoths, which had been treated with the afu UDG enzyme leaving post-mortem DNA damage at the ends of the molecules (Supplementary Tables 2 and 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 v0.7.8 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 v1.10[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, Chukochya), we took several steps to assess the authenticity and quality of the data (Supplementary Section 4). First, only reads that mapped uniquely to non-repetitive regions of the LoxAfr4 reference and had a mapping quality ≧30 were retained, whereas 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 employed 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 20-35 bp and at 5 bp intervals between 35-50 bp (Supplementary Section 4). This allowed us to identify a sample-specific minimum read length cutoff, above which we consider reads to be correctly mapped and endogenous (Supplementary Section 4, Supplementary Table 3). Based on this, we applied the longest sample-specific cutoff (≥35 bp, Krestovka) for all samples. We used mapDamage v2.0.6[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 (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 v0.921[36]. For base calling we only considered 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 mtDNA 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 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 37.8x, 47.5x, and 77.1x for Adycha, Krestovka, and Chukochya, and 99.6x and 179.5x for Scotland and Kanchalan, respectively. These assemblies were then aligned using Muscle v3.8.31[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 v2.1.10[42] (Supplementary Section 5). To estimate the age of the three oldest *Mammuthus* samples (Adycha, Krestovka, Chukochya), we performed a Bayesian reconstruction of the phylogenetic tree using BEAST v1.10.4[43]. We calibrated the molecular clock using tip ages for all ancient samples with a finite radiocarbon date, as well as a lognormal 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] (see Supplementary 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 lognormal distributions with priors based on stratigraphic context and previous genetic studies, using two MCMC chains of 100 million generations, sampling every 10,000 and discarding the first 10% as burn-in (Supplementary Table 5, Supplementary Section 5).

Genetic dating based on autosomal data

Specimen age estimates for Adycha and Chukochya (Krestovka was excluded as too few autosomal bases were available for this analysis) were estimated based on the autosomal data following the method described in Meyer *et al.*[17], using the American mastodon (*Mammut americanum*), which is 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 elephant reference genome by requiring that the alignments of the mastodon, two African 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 (nW − nM)/nW, based on the number of derived changes in the Wrangel genome (nW) and the other *Mammuthus* genome (nM), using an assumed divergence time of 5.3 million years[15] to the common ancestor of African elephant and woolly mammoth. Age variance estimates were calculated in windows of 5 Mb and we computed bootstrap confidence intervals as 1.96× standard error around the date estimates (Supplementary Section 6).

Nuclear genetic relationships and phylogeny

We reconstructed phylogenetic trees based on the whole genome Identical-By-State (IBS) 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 (ME) method as implemented in FASTME[45] (Fig. 1b, Supplementary Section 7). Next, we inferred relative population split times using an approach that examines single nucleotide polymorphic (SNP) 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 SNPs with a quality ≥30, and filtered out all SNP in repetitive regions, within 5 bp from indels, at CpG sites and sites below 1/3 or above two times 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, f4 statistics, AdmixtureGraphs and TreeMix

We first used Admixtools v5[22] to calculate D- and f_4-statistics for all possible quadruple combinations of samples iterating through the three different groups (P_1, P_2, P_3,) based on the randomly sampled alleles, conditioning on all sites that are polymorphic among the 6 Asian elephant genomes[22]. The mastodon was used as an outgroup in all comparisons (Supplementary Table 6, 7). Direct estimates of genomic ancestries using f_4-ratios were additionally calculated for specific pairs in AdmixTools (Supplementary section 9)[22]. 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 two admixture events) relating these three individuals, one woolly mammoth (Wrangel), one Columbian mammoth, and one Asian elephant, setting the latter as outgroup (Supplementary Section 8). We repeated the admixturegraph analysis using the above described f_4-statistic with qpBrute[46], which in addition allowed 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. Where a node could not be inserted without producing f_4 outliers (i.e. $|Z| >=3$), 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 v1.12[21]. We first estimated the allele frequencies among the randomly sampled alleles and subsequently ran the TreeMix model accounting for linkage disequilibrium (LD) by grouping sites in blocks of 1,000 SNPs (-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 based on the distribution of private sites. Briefly, we estimated the number of callable sites, the SNP 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 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 the ‘Nuclear genetic relationships and phylogeny’ section.

Genetic adaptations of the woolly mammoth

To investigate the timing of genetic adaptations in the woolly mammoth lineage, we used *last* v1170[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 (last.cbrc.jp). The Picard Liftover tool (‘Picard Toolkit’, 2019) 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 where 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, Chukochya) and the Columbian mammoth (Supplementary Table 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 v1.3.1[50] to identify genes that potentially have been under positive selection in Late Pleistocene woolly mammoths (Supplementary Table 11, Supplementary Section 10).

Extended Data figure legends

Extended Data Fig. 1. Mammoth molars and morphometric comparisons. a-b, upper third molars in lateral and cross-sectional views; **c**, partial lower third molar in lateral and occlusal views. **a**, Chukochya (PIN-3341-737); **b**, Krestovka (PIN-3491-3) flipped horizontally; **c**, Adycha (PIN-3723-511), occlusal view flipped horizontally. Note the more closely-spaced lamellae and thinner enamel in **a** (*primigenius*-like) than **b** and **c** (*trogontherii*-like). **d**, Hypsodonty index vs lamellar length index of upper M3s; **e**, Enamel thickness index vs basal lamellar length index of lower M3s. Olyorian specimens yielding DNA are labelled by site name. Green dashed line: convex hull summarising Early to early Middle Pleistocene (ca. 1.5-0.5 Ma) North American *Mammuthus* samples (data points not shown). Green and blue squares: Early and Late Olyorian North-East Siberian samples, respectively; red and green circles: European *M. meridionalis* and *M. trogontherii*, respectively; blue circles, *M. primigenius* from North-East Siberia and Alaska. Note (i) similarity of Krestovka and Adycha to other Early Olyorian molars and to European steppe mammoths (*M. trogontherii*), (ii) similarity of early North American mammoths to these (Early Olyorian in particular), (iii) similarity of Chukochya to *M. primigenius*. For site details, measurement definitions and data, see Supplementary Section 1.

Extended Data Fig. 2. Sample age based on biostratigraphy, paleomagnetic reversals and genomic data. Chart shows the stratigraphic position of the Kutuyakhian fauna, *Phenacomys* complex, Early Olyorian and Late Olyorian faunas in relation to important European, northwest Asian and northern North American stratigraphic benchmarks. ELMA - European Land Mammal Ages (small mammals), LMA - Land Mammal Ages (large mammals), MN/MQ - European Small Mammal Biozones, EEBU – East European biochronological units. Biostratigraphic and palaeomagnetic based chronological constraints for the specimens are provided, in comparison with the DNA-based age estimations.

Extended Data Fig. 3. DNA fragment length distributions for nine mammoths. Reads are aligned to the LoxAfr4 autosomes. For the three Early-Middle Pleistocene samples (Krestovka, Adycha, Chukochya), reads of 25-200 bp length are shown, whereas 30-200 bp reads are shown for the remaining samples. Ultrashort reads (<35 bp) are denoted in red and were shown to be enriched for spurious alignments and therefore excluded from downstream analyses (Supplementary Section 4). The mean read lengths (μ) were calculated using only the retained reads (≥35 bp).

Extended Data Fig. 4. Post-mortem cytosine deamination damage profiles at CpG sites. The most ancient samples (Krestovka, Adycha, Chukochya) carry a greater frequency of cytosine deamination compared to younger permafrost preserved woolly mammoth samples (Oimyakon and Wrangel) and the Columbian mammoth (*M. columbi*) specimen.

Extended Data Fig. 5. F(A|B) statistics. The statistics reflect relative divergence between the genomes on the left and the right side. Lower values indicate reduced derived allele sharing between the sample indicated on the left and the right of the graph, at sites for which the genome on the right panel is heterozygous. The lower the value, the more drift has occurred between the genomes and thus the older their genetic divergence.

Extended Data Fig. 6. qpGraph model. The most parsimonious graph model (highest Bayes Factor) of the phylogenetic relationships among mammoths lineages augmented with one admixture event. Branch lengths are given in f-statistic units multiplied by 1,000. Discontinuous lines show admixture events between lineages, with percentages representing admixture proportions.

Extended Data Fig. 7. Ghost introgression analysis of the Columbian mammoth genome. **a,** The number of private alleles per 1000 bp within genomic regions identified as woolly mammoth (*M. primigenius*) ancestry or ghost ancestry. **b,** Maximum-likelihood phylogenies for those genomic regions identified as ghost ancestry in the Colombian mammoth (*M. columbi*) genome. **c,** Maximum-likelihood phylogenies for regions identified as un-admixed ancestry.

Acknowledgments

T.v.d.V, P.P. and D.D.d.M., M.D. and L.D. acknowledge support from the Swedish Research Council (2012-3869 & 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. MH, JAT and GX were supported by NERC (grant no. NE/J009490/1) and the ERC StG grant GeneFlow (#310763). B.S. and J.O. were supported by the U.S. 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. Neil Clark at the Hunterian Museum kindly provided access to the Scotland mammoth sample. Finally, we wish to especially acknowledge the seminal work of our late friend and colleague Andrei Sher, who in many years of fieldwork defined and described the Olyorian sequence, collected large quantities of fossil vertebrate material including all the Early/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/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- and f4-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 interpret the results. Morphological analyses as well as palaeontological and geological information was provided by P.N. and A.M.L. The manuscript was written by T.v.d.V., P.P., D.D.d.M., P.N. and L.D., with contributions from all coauthors.

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.com/stefaniehartmann/readLengthCutoff.

Competing Interests

The authors declare no competing interests.

Additional Information

Supplementary information is available for this paper at https://doi.orgxxxxx.

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

References (Methods)

28. Gansauge, M.-T. & Meyer, M. Single-stranded DNA library preparation for the sequencing of ancient or damaged DNA. *Nat. Protoc.* **8**, 737–748 (2013).

29. John, J. S. SeqPrep: Tool for stripping adaptors and/or merging paired reads with overlap into single reads. *URL: https://githubcom/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. *arXiv [q-bio.GN]* (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). https://doi.org/10.1186/s12864-020-07229-y

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. U. S. A.* **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. (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**, (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).

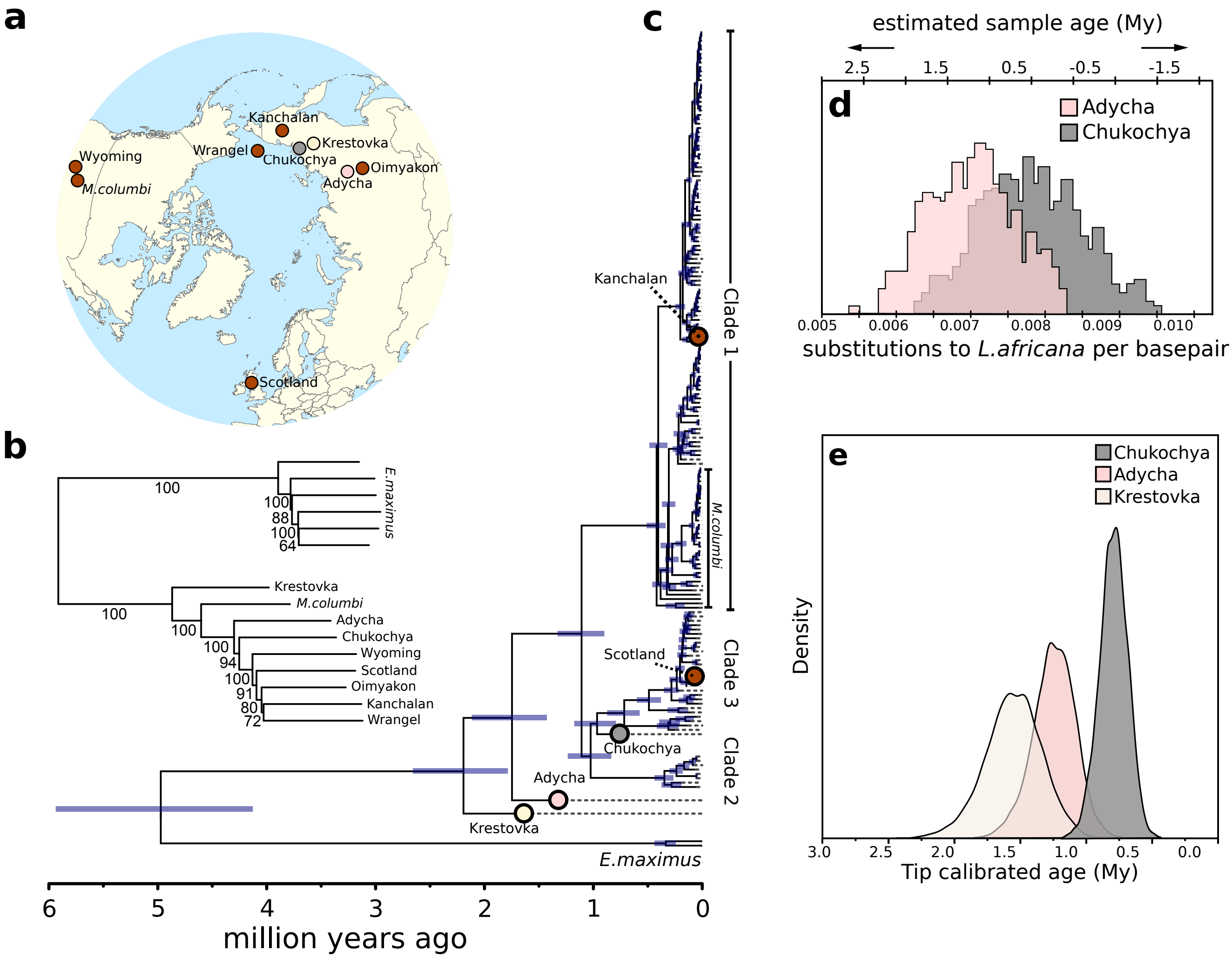
46. Liu, L. *et al.* Genomic analysis on pygmy hog reveals extensive interbreeding during wild boar expansion. *Nat. Commun.* **10**, 1992 (2019).

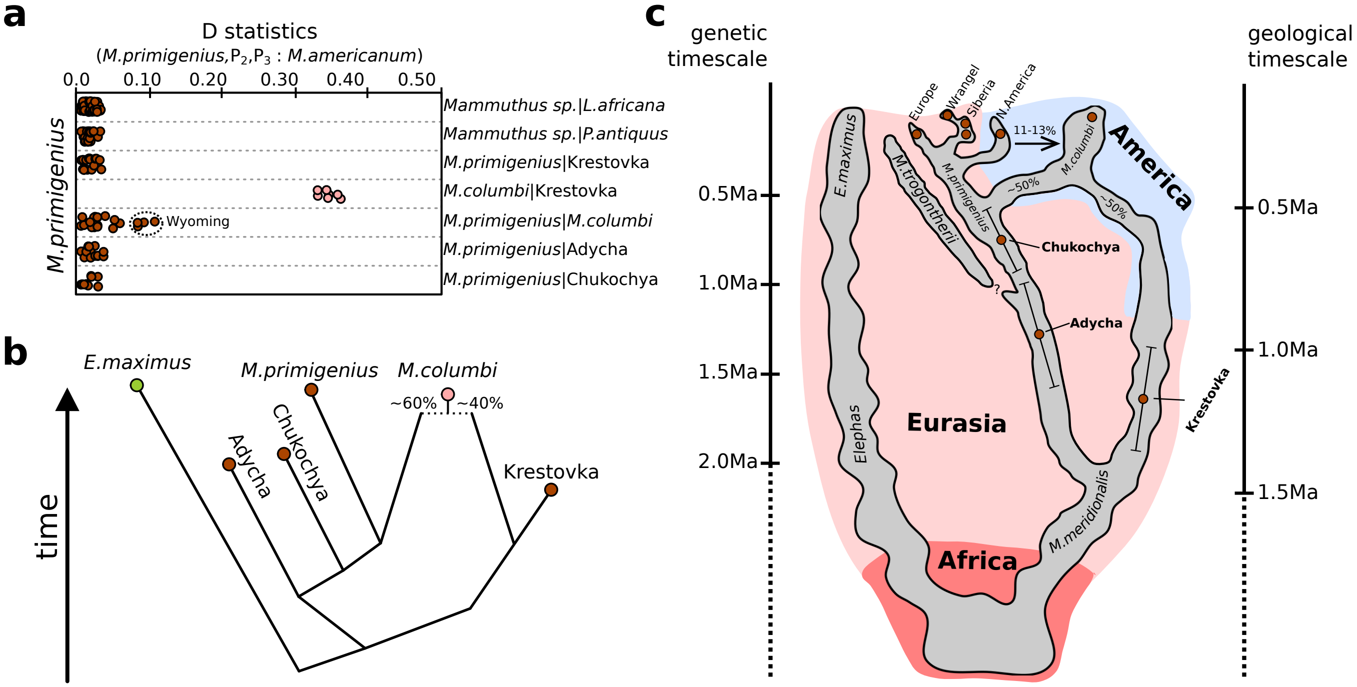
47. Frith, M. C., Hamada, M. & Horton, P. Parameters for accurate genome alignment. *BMC Bioinformatics* **11**, 80 (2010).

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).


a
Kanchalan
Krestovka
Wrangel
Chukochya
Wyoming
Oimyakon
Adycha
M.columbi
Scotland
b
100
E.maximus
100
88
100
64
Krestovka
M.columbi
100
Adycha
100
Chukochya
100
Wyoming
94
Scotland
100
Oimyakon
91
Kanchalan
80
Wrangel
72
c
Kanchalan
Clade 1
M.columbi
Scotland
Clade 3
Chukochya
Clade 2
Adycha
Krestovka
E.maximus
6
5
4
3
2
1
0
million years ago
estimated sample age (My)
2.5
1.5
0.5
-0.5
-1.5
d
Adycha
Chukochya
0.005
0.006
0.007
0.008
0.009
0.010
substitutions to L.africana per basepair
e
Chukochya
Adycha
Krestovka
Density
3.0
2.5
2.0
1.5
1.0
0.5
0.0
Tip calibrated age (My)


a
D statistics
(M.primigenius,P2,P3 : M.americanum)
0.0
0.10
0.20
0.30
0.40
0.50
M.primigenius
Mammuthus sp.|L.africana
Mammuthus sp.|P.antiquus
M.primigenius|Krestovka
M.columbi|Krestovka
M.primigenius|M.columbi
M.primigenius|Adycha
M.primigenius|Chukochya
Wyoming
b
time
E.maximus
M.primigenius
M.columbi
~60%
~40%
Adycha
Chukochya
Krestovka
c
genetic timescale
0.5Ma
1.0Ma
1.5Ma
2.0Ma
geological timescale
0.5Ma
1.0Ma
1.5Ma
Europe
Wrangel
Siberia
N.America
11-13%
America
E.maximus
M.trogontherii
M.primigenius
M.columbi
~50%
~50%
Chukochya
?
Adycha
Krestovka
Eurasia
Elephas
M.meridionalis
Africa