Late Quaternary dynamics of Arctic biota from ancient environmental genomics

https://doi.org/10.1038/s41586-021-04016-x

Received: 21 March 2021

Accepted: 13 September 2021

Published online: 20 October 2021

Open access

Check for updates

A list of authors and affiliations appears at the end of the paper.

During the last glacial–interglacial cycle, Arctic biotas experienced substantial climatic changes, yet the nature, extent and rate of their responses are not fully understood[1–8]. Here we report a large-scale environmental DNA metagenomic study of ancient plant and mammal communities, analysing 535 permafrost and lake sediment samples from across the Arctic spanning the past 50,000 years. Furthermore, we present 1,541 contemporary plant genome assemblies that were generated as reference sequences. Our study provides several insights into the long-term dynamics of the Arctic biota at the circumpolar and regional scales. Our key findings include: (1) a relatively homogeneous steppe–tundra flora dominated the Arctic during the Last Glacial Maximum, followed by regional divergence of vegetation during the Holocene epoch; (2) certain grazing animals consistently co-occurred in space and time; (3) humans appear to have been a minor factor in driving animal distributions; (4) higher effective precipitation, as well as an increase in the proportion of wetland plants, show negative effects on animal diversity; (5) the persistence of the steppe–tundra vegetation in northern Siberia enabled the late survival of several now-extinct megafauna species, including the woolly mammoth until 3.9 ± 0.2 thousand years ago (ka) and the woolly rhinoceros until 9.8 ± 0.2 ka; and (6) phylogenetic analysis of mammoth environmental DNA reveals a previously unsampled mitochondrial lineage. Our findings highlight the power of ancient environmental metagenomics analyses to advance understanding of population histories and long-term ecological dynamics.

Climate changes are amplified at high latitudes and have pronounced effects on Arctic ecosystems[1]. Their effects on Arctic plant and animal communities, as well as the human populations who are dependent on them, would have been especially pronounced during the extremely cold and arid Last Glacial Maximum (LGM) (26.5–19 ka)[2] and later during the rapid warming that preceded the Holocene. However, precisely what those effects were, and how they played out across the Arctic, are not fully understood. These dynamics were further complicated by differences in the timing and extent of glaciation in different regions across this vast and topographically complex landscape. Previous studies based on pollen and plant macrofossils have documented substantial spatiotemporal variations in Arctic vegetation over the past 50,000 years (50 kyr)[1,3], yet it continues to be debated how climatic changes during this period affected plant communities in different regions of the Arctic, and how changes in climate and vegetation may have affected large mammals (that is, megafauna)[4–6]. Skeletal remains show that several megafaunal species, including woolly mammoth (*Mammuthus primigenius*), woolly rhinoceros (*Coelodonta antiquitatis*), steppe bison (*Bison priscus*) and horse (*Equus* spp.), were abundant in the Arctic during the Pleistocene epoch, but are thought to have become regionally or globally extinct by the onset of the Holocene[4,5]. However, the precise timing of megafaunal extinctions, and whether and to what extent some of these taxa survived into the Holocene, is uncertain. Similarly, the contribution of various abiotic and biotic drivers to the extinction process of different taxa remains an open question[7,8].

To address these knowledge gaps, we performed a metagenomics analysis of ancient environmental DNA (eDNA) of plants and animals recovered from sediments from sites distributed across much of the Arctic covering the past 50 kyr. Relative to other palaeoecological proxies (such as pollen and macrofossils), ancient eDNA offers distinct advantages—including greater taxonomic resolution across the full tree of life[9] and higher spatial and temporal precision than pollen—as eDNA mainly derives from the local community[10]. We used metagenomic analysis rather than the widely used metabarcoding approach because it enables the sequencing of DNA fragments from entire genomes without taxon-specific amplifications, therefore improving the specificity and sensitivity of taxonomic identification, as well as facilitating the authentication of endogenous ancient DNA from modern contaminants[9]. However, metagenomic analysis requires genome-scale reference data, which are limited for most regions of the world, including the Arctic. Thus, a key component of our study is the generation of a substantial corpus of plant reference sequences.

Metagenomic dataset and database

We generated the eDNA metagenomic dataset from 535 sediment samples obtained at 74 circumpolar sites (Fig. 1). Samples come from lake sediments and stratigraphic exposures (unconsolidated permafrost). For the purpose of understanding regional variability, we grouped sites into four regions: North Atlantic; northwest and central Siberia; northeast Siberia; and North America (Fig. 1). Sample ages span the past 50 kyr, albeit in varying numbers, from all regions with the notable exception of the North Atlantic, which was largely covered by ice sheets that often erased pre-LGM deposits[2,11].

From the 535 samples, we generated 10.2 billion sequencing reads that passed the filtering criteria and were used for analysis (Methods). We created a comprehensive reference database for taxonomic identification by merging the NCBI-nt and NCBI-RefSeq databases, and supplemented the limited genomic-scale public reference data for Arctic species with 12 Arctic animals and an extensive sequencing effort of 1,541 modern Holarctic plant genome skims (PhyloNorway; Methods). These new sequences comprise 311.3 million whole-genome contigs and provide a broader and more reliable plant reference database than previously available. The merged reference database contains a total of 380.4 million entries and covers about 1.47 million organisms. We developed a *k*-mer-based method to evaluate the availability and coverage of our combined reference database for different taxa (Methods) and found that it covers a wide range of both Arctic and non-Arctic species (Supplementary Information 9.2.3). Accordingly, the addition of our new reference genomes did not cause bias towards Arctic taxa, providing confidence in our identifications. We used robust approaches to identify taxa from individual reads and collated the resulting taxonomic composition at the generic or familial level (Methods). We applied several methods to authenticate the plant and animal taxonomic profiles; the identifications were reliably classified despite the short DNA sequences that were preserved in these samples (Methods).

Moreover, 131 samples in this dataset were processed for metabarcoding, targeting the short DNA barcodes of plants[12], enabling a comparison between the two approaches (Methods). The results showed that the metagenomic analysis captured greater floristic and faunal diversity and achieved better taxonomic resolution (Supplementary Information 11.2). We also found that only about 1.26% of the plant DNA reads are of ribosomal and chloroplast origin (Supplementary Information 9.2.5), suggesting that the metabarcoding approach—which relies on organelle DNA—makes use of only a small fraction of preserved DNA. However, we acknowledge that these comparisons are sample- and method-specific; more studies are needed before broader conclusions about the relative merits of the two approaches can be reached.

Circum-Arctic vegetation dynamics

We combined plant assemblages that were reconstructed from all of the samples to describe the temporal changes in floristic composition, diversity and community structure across the Arctic (Fig. 2a and Extended Data Fig. 1). Our results show substantial and repeated responses of Arctic vegetation to changing climates over the past 50 kyr.

The overall floristic diversity increased steadily from 50 ka and reached its highest levels at the onset of the LGM (about 26.5 ka), when the climate reached its coldest and driest point at many locations[2,11] (Fig. 2a). Vegetation turnover was high before about 38 ka, and the identified shrubs, forbs and grasses suggest a shifting mosaic of steppe–tundra vegetation. Herbaceous plants were the dominant plant group until about 19 ka, with forbs more abundant than graminoids (Fig. 2a), but not as dominant as suggested by a previous metabarcoding study[12]. Trees and aquatic plants were limited in distribution to lower-latitude sites—consistent with overall dry and cold climate conditions during this period. The scarcity of cold-tolerant trees such as *Pinus* and *Picea*, and absence of *Larix*, reflect low precipitation and strong winds (Fig. 2a and Extended Data Fig. 1a).

The transition into the LGM featured declining temperature and precipitation (Fig. 2a). Across the Arctic, trees remained absent, and there was a sharp decrease in floristic diversity, mainly caused by the decline in herbaceous taxa. Overall, vegetation turnover was consistently high during this decline in diversity, suggesting that cold and dry extremes caused the loss of taxa from all plant communities, although the taxa that were dominant in the pre-LGM period remained (Extended Data Fig. 1a). LGM vegetation dissimilarity was the lowest of all time periods (Extended Data Fig. 1b, c), indicating considerable homogeneity across much of the unglaciated Arctic.

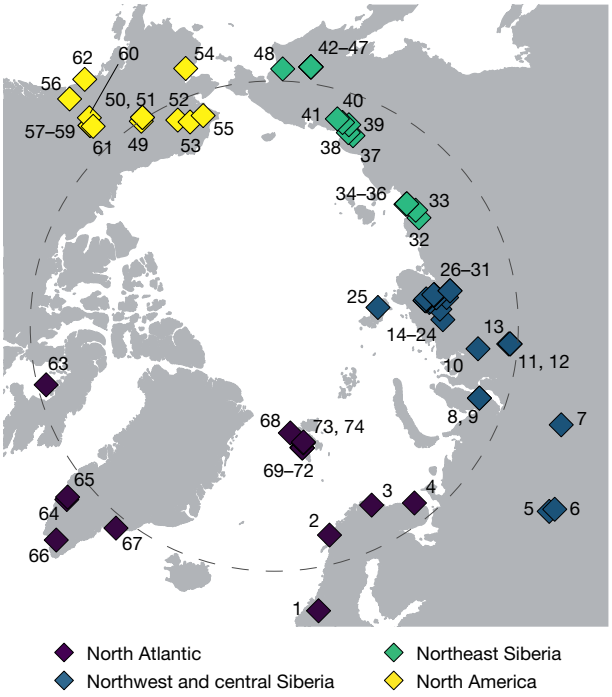


Fig. 1 | Site distribution (North Pole-centred view). Samples (n = 535) from a total of 74 circumpolar sites were grouped into four geographical regions (Supplementary Information 2). The grey dashed circle indicates the Arctic Circle (66.5° N). Site IDs are labelled on the map. The corresponding information is provided in Supplementary Data 1.

After the LGM, warming towards the Bølling–Allerød interstadial (approximately 14.6–12.9 ka)[13] led to vegetation divergence among sites (Extended Data Fig. 1b, c). There was a substantial increase in the abundance of woody plants (such as *Salix* and *Betula*), whereas the herbaceous diversity continued to decline, causing the overall diversity to reach its lowest point at the beginning of the cold Younger Dryas stadial (approximately 12.9–11.7 ka)[14] (Fig. 2a). The abundance of woody taxa and vegetation turnover rate reached the highest point during the Younger Dryas; the latter is consistent with the intensive climate changes that mark the transition from the Pleistocene to the Holocene.

Shortly after the Younger Dryas, summer insolation peaked and atmospheric CO_2 reached Holocene levels[15]. Previously abundant plant taxa such as *Artemisia* and *Poa* rapidly declined or vanished locally. Other plant taxa, particularly boreal trees and prostrate shrubs (such as *Vaccinium*), appeared and later became abundant (Extended Data Fig. 1a), suggesting that there was a shift from open, cold-adapted tundra–steppe to a mosaic of herbaceous and woody plant communities. The floristic diversity of this more mesophilic vegetation increased during the Early Holocene as climate continued to warm and effective precipitation increased, but then declined during the middle Holocene (Fig. 2a).

Owing to dating uncertainties and limits on the temporal resolution of palaeoclimatic simulations, our results captured only broader changes in vegetation dynamics under climate change. During much of the past 50 kyr, overall plant diversity decreased when the proportion of trees and shrubs increased, as they outcompete herbaceous taxa through shading[16]. By contrast, when climate became more suitable for herbaceous taxa, diverse taxa expanded to share the landscape, and the overall diversity therefore increased.

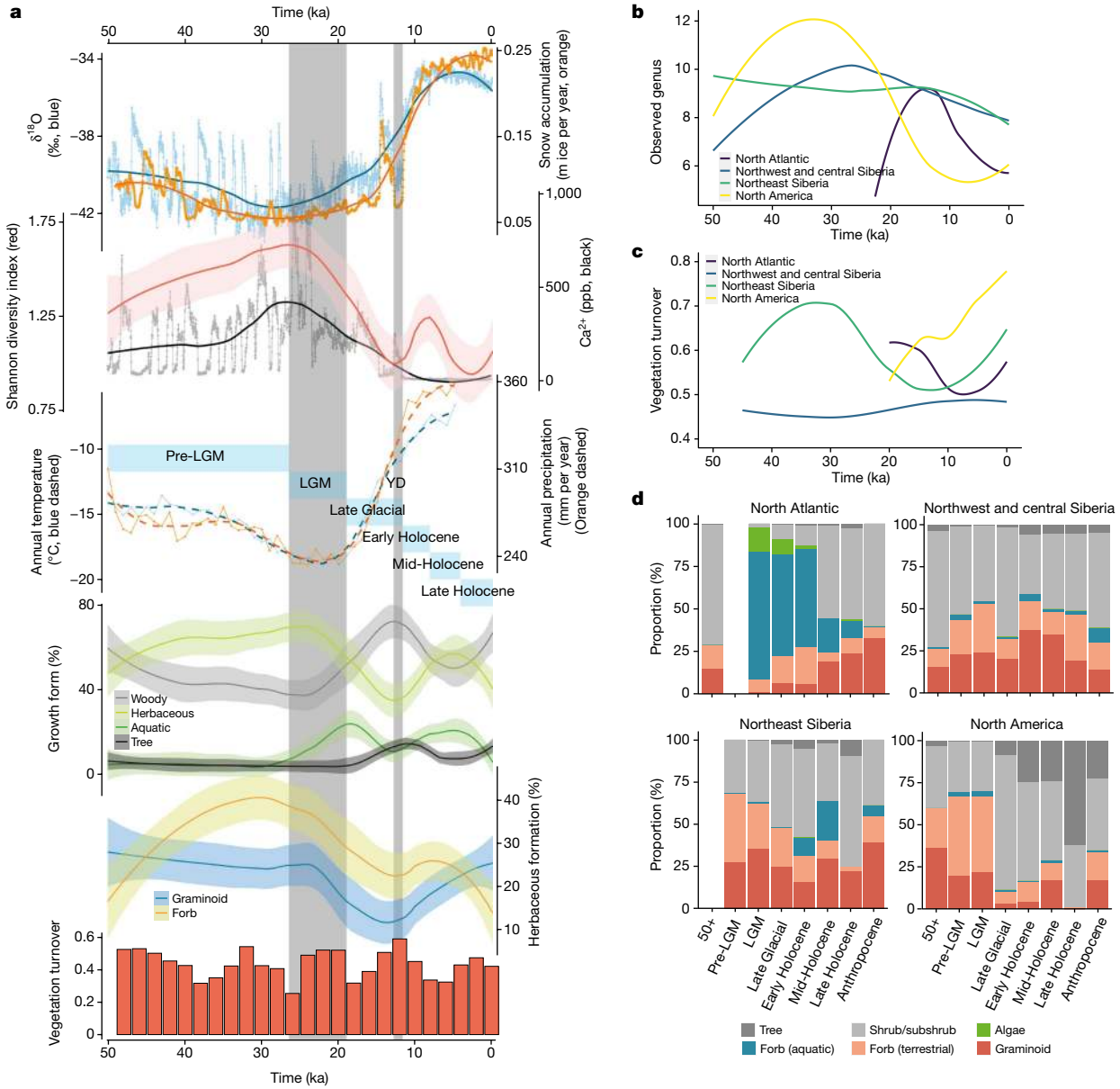


Fig. 2 | Climate and vegetation changes over the past 50 kyr. a, Pan-Arctic climate changes and vegetation variations. LGM (26.5–19 ka) and Younger Dryas (YD) (12.9–11.7 ka) are indicated by grey bars. The six time intervals are indicated by light blue bars (Supplementary Information 2). The error bands denote s.e. From top to bottom (see Methods for detailed calculations): the Greenlandic ice-core $\delta^{18}O$ ratio and snow accumulation rate; the plant Shannon diversity and the Greenlandic ice-core calcium concentration; the average modelled annual temperature and precipitation for all eDNA sampling sites; the proportion of plant growth forms; the proportion of the herbaceous plant growth forms; and the vegetation turnover rates. **b**, The number of observed genera in different regions. **c**, Regional vegetation turnovers. **d**, Regional vegetation morphological compositions. The sample sizes for each region and time interval are provided in Supplementary Information 2. Calculations are supplied in the Methods.

Regional vegetation dynamics

Underlying the generalized pattern of Holarctic vegetation changes are significant geographical differences. Early in postglacial times, the North Atlantic experienced the sharpest rises in taxonomic richness (Fig. 2b), along with the steepest temperature increase (Extended Data Fig. 2b). The increase in postglacial richness was probably driven by species dispersals coupled with habitat diversification[17], that is, gynomorphically dynamic substrates that were exposed by glacial retreat and shaped by meltwater. The resultant vegetation initially had low diversity but was rich in aquatic taxa (Fig. 2b, d). The abundance of aquatic taxa relates in part to the prevalence of samples from lakes in the North Atlantic (Supplementary Information 10), but nonetheless highlights the ability of aquatic plants to disperse rapidly into newly deglaciated terrain containing abundant streams and lake basins[18]. As the postglacial climate continued to warm, the overall proportion of aquatic taxa declined as trees and shrubs (for example, *Betula*, *Salix* and *Vaccinium*) became abundant in this region (Fig. 2d and Extended Data Fig. 3).

Northeast Siberia and North America experienced less radical postglacial changes in vegetation type (Fig. 2c, d). During the Late Glacial, trees and shrubs became more widely distributed, and floristic diversity started to decline—a trend that was especially pronounced in North America (Fig. 2b, d). By about 12 ka, rising sea levels had flooded the Bering Strait, and the vegetation on each side started to diverge (Extended Data Fig. 2a). In northeast Siberia, greater effective moisture

within the Holocene led to the expansion of aquatic plants (such as *Hippuris* and *Menyanthes*). The previously dominant steppe taxa (for example, *Poa* and *Artemisia*) declined, although sedges, of which many species are hygrophilous, continued to be abundant (Extended Data Fig. 3). The vegetation of this region became a mosaic of steppe and tundra elements. In North America, trees such as *Populus* and *Picea* became more widespread during the Early Holocene and previously widespread steppe species declined (Fig. 2d and Extended Data Fig. 3). A broad, southern swath of eastern Beringia became boreal forest.

In contrast to the changes observed in these regions, vegetation in northwest and central Siberia remained relatively unchanged through the Pleistocene–Holocene transition (Fig. 2c, d). However, some cold- and/or dry-adapted taxa (such as *Artemisia* and *Poa*) were replaced by forbs that were better adapted to warmer climates, and *Salix* was partially replaced by *Betula* and *Alnus* (Extended Data Fig. 3). The vegetation in this region persisted as a steppe–tundra mosaic through much of the Holocene, probably due to central Siberia's extreme climatic continentality caused by the Siberian anticyclone[19], which created largely ice-free conditions during the LGM and fostered dry hydrogeological conditions in postglacial times that mitigated the effects of rising global temperatures on vegetation[11].

Overall, these results show that postglacial plant communities regionally diverged in response to warming temperatures, increasing moisture, retreating ice sheets and marine transgressions. Although regions that were once overridden by continental ice sheets experienced extreme vegetation changes, the vegetation in unglaciated interior regions remained rather stable. This maritime–continental contrast highlights the importance of moisture in driving ecosystem changes in the Arctic[7,20]. We next incorporate these insights into vegetation dynamics, together with other potential drivers, into a model to identify the factors influencing animal distributions.

Animal distribution drivers

We developed a model using reconstructed animal distributions and floristic compositions, modelled palaeoclimate variables and inferred human occurrences (Methods) to examine the relative effects of abiotic and biotic factors on Arctic mammal distributions over the past 50 kyr.

We found that certain herbivores tend to co-occur in time and space. For example, the eDNA presences of caribou, hare and vole are statistically strong co-indicators for the presence of horse and mammoth eDNA (Fig. 3). This suggests that co-existence was more common among Arctic herbivores than interspecies exclusion[21]. By contrast, the distribution of humans over time was almost entirely unrelated to the presence of most herbivores (apart from hares) (Fig. 3). Given that the model purposefully overestimated the presence of humans (Methods), their largely independent distributions from megafauna, their sparseness in the high Arctic before 4 ka (Supplementary Data 7) and the scarcity of kill sites in archaeological records, the notion of human overkill as the cause of Arctic megafaunal extinction is highly improbable[6,8]. Interestingly, the only predator–prey relationship of note in the model is the significant positive effect of caribou on the distribution of wolves (Fig. 3), probably reflecting that the wolf is well-adapted to hunt caribou.

To better gauge the explanatory power of environmental variables, we removed the effects of the presence of the eDNA of other animals (Extended Data Fig. 4a and Methods). The most consistent and widely prevalent patterns are the generally negative effects of plant NMDS1 and NMDS3—the first and third components of the non-metric multidimensional scaling (NMDS) of the vegetation compositions (Methods)—on the presence of animal eDNA. Plant NMDS1 reflects an aquatic-to-terrestrial plant gradient, and plant NMDS3 reflects a graminoids-to-woody plant gradient, particularly sedges within the graminoids, which include species that are prominent in present-day wetland communities (Extended Data Fig. 4b). These two negative covariates apply to the distribution of both small (vole and hare) and large (horse and mammoth) mammals, indicating that a wetter environment with a high proportion of hygrophilous plants (that is, moisture-loving plants) was a key factor restricting animal distributions. The distribution of mammoths tends to be positively affected by plant NMDS2, which mainly reflects the proportion of woody plants (particularly shrubs and subshrubs) as opposed to herbaceous plants, whereas the reverse is true for horses (Fig. 3). We also found that horses are more sensitive to vegetation composition compared with other herbivores (Supplementary Information 13.3). These findings support the hypothesis that horses were more restricted to a grassland environment and may also indicate a greater dietary flexibility in mammoths.

When each herbivore species is considered individually, the only climate variable that is consistently and positively associated with the presence of their eDNA is temperature seasonality (Fig. 3 and Extended Data Fig. 4a), consistent with expectations based on the continental climate associated with the Mammoth Steppe, a biome that is associated with extremely cold and dry conditions that supported abundant large mammal grazers[19]. The importance of climatic variables becomes more evident when herbivores are considered as a group. Precipitation—in greater amounts and seasonality—is a principal negative factor in the distribution of Arctic herbivores (Fig. 3), presumably because increased snow cover during winter limited the food access of grazers, and a wetter substrate is more difficult for them to exploit, in contrast to the firm and dry ground of the steppe–tundra[7,19].

Late-surviving megafauna

The timing of Arctic megafaunal extinction is a matter of debate, not least because last appearance dates (LADs) are repeatedly revised as younger fossils are reported[5,6], and also because discovering the remains of the last surviving individuals of a species is extremely unlikely[22]. As a result, LADs systematically underestimate when a species disappeared, raising the possibility that populations persisted longer than is now evident[4,23]. The extinction timing can be better gauged with eDNA; an animal leaves behind only a single skeleton, which is much less likely to be preserved, recovered and dated, when compared with the amount of DNA it continuously spread into the environment while it was alive.

Our data indicate that mammoths survived into the Early Holocene in present-day continental northeast Siberia until 7.3 ± 0.2 ka (seven samples younger than 10 ka) and North America until 8.6 ± 0.3 ka. Notably, we recovered mammoth DNA from a series of samples from the Taimyr Peninsula that indicate the presence of mammoths in north central Siberia as late as 3.9 ± 0.2 ka (site LUR10) (Fig. 4 and Supplementary Information 3.3). The survival of mammoths into the Holocene in these regions is probably attributable to the persistence of the steppe–tundra vegetation of dry- and cold-adapted herbaceous plants that was present during the Pleistocene (Fig. 2d). This vegetation would have provided a suitable habitat for mammoths and possibly other dryland grazers such as horses (Extended Data Fig. 5), which are known to have survived in the region until at least 5 ka (ref. [24]). Together, these eDNA results indicate that mammoths survived much longer than previously thought—which, on the basis of skeletal remains, was around 10.7 ka on continental Eurasia[25] and around 13.8 ka in Alaska[8]. Given that humans occupied northern Eurasia sporadically from at least 40 ka and continuously after 16 ka (refs. [26,27]), the late-surviving Taimyr mammoths potentially encountered and co-existed with humans over at least a 20-kyr interval, therefore giving no support to the human overkill (blitzkrieg) model that postulates the mammoth extinction occurred within centuries after the first human contact[6].

We also detected woolly rhinoceros DNA as late as 9.8 ± 0.2 ka in northeast Kolyma, horse DNA in Alaska and the Yukon as late as 7.9 ± 0.2 ka, and bison as late as 6.4 ± 0.6 ka in high-latitude localities of northeast Siberia (Extended Data Fig. 5). All of these instances represent substantially later LADs than fossil-based dates (that is,

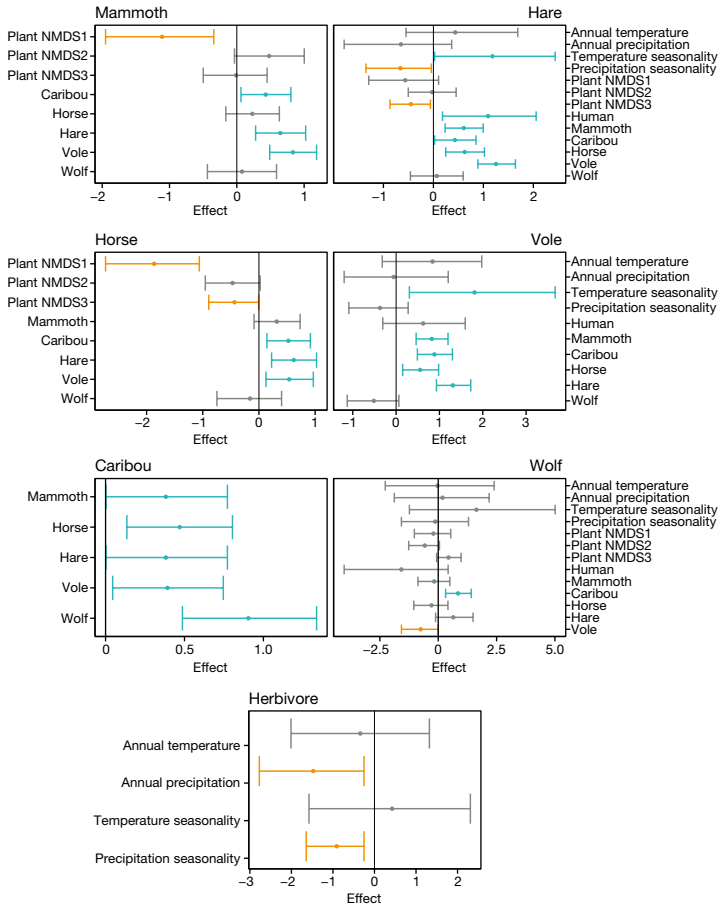


Fig. 3 | Spatiotemporal models to retrodict the explanatory factors for animal distribution. The values indicate posterior parameter estimates of covariate effects for the models explaining the presence–absence of each animal's eDNA. Only covariates included in the model with lowest Watanabe–Aikaike information criterion are shown (Methods). The dots represent the posterior means, and the whiskers represent the posterior 2.5% and 97.5% quantiles. Covariate effects of which the 2.5% and 97.5% quantiles are both negative (red), and effects of which the 2.5% and 97.5% quantiles are both positive (blue) are indicated.

for woolly rhinoceros in Eurasia, about 14 ka (ref. [28]); and for horses and steppe bison in Alaska, 12.5 ka (refs. [5,8]). Collectively, these findings highlight the value of eDNA in improving megafauna extinction chronologies.

Population diversity of megafauna

Megafaunal eDNA from across the Arctic also enables us to resolve population-level patterns, which is crucial for uncovering species-specific demographic and evolutionary responses to past climatic and environmental changes. We applied a method for phylogenetically assigning the identified eDNA to mitochondrial haplogroups of mammoth and horse, the two most abundant species detected in our dataset (Methods).

A mammoth phylogeny composed of four previously described major mitochondrial clades (clade 1, including 1C and 1DE, and clades 2 and 3)[29] was reconstructed from 78 mammoth mitochondrial genomes. The recovered mammoth eDNA was then assigned to a best-fit node on the tree based on single-nucleotide polymorphism (SNP) support/conflict, enabling clade assignment for 79 eDNA samples (Extended Data Fig. 6).

The mammoth haplogroups that we identified are consistent with those that were previously identified from fossil remains and have comparable biogeographical and biostratigraphic distributions (Fig. 4). Overall, clade 3 was present mainly in Europe and northwest Siberia, whereas clade 2 occurred mostly in central and northeast Siberia. Clade 1 was widely scattered across North America and the Asian Arctic, with 1DE occurring throughout Siberia and 1C in North America. Temporally, clades 2 and 3 were the older lineages, and disappeared between 40 ka and 30 ka. Only clade 1 survived past the LGM, with the last 1C individual dating to 10.35 ka. Like the late-surviving mammoths on Wrangel Island[30], the late-surviving mammoths on mainland Siberia were also members of 1DE, the only clade detected to date that postdates the Early Holocene (that is, after 8.2 ka). However, despite belonging to the same clade, none of the mainland late-surviving populations is placed in the Wrangel Island haplogroup (Extended Data Fig. 6). Furthermore, we note that two mammoth eDNA samples (cr5_11 and tm4_13) attach to the existing tree at the shared root of clades 2 and 3 (Extended Data Fig. 6), with cr5_11 containing many sequence variants not found in previously sequenced samples (Supplementary Information 14.1.2), suggesting that they represent a separate and previously unrecorded mitochondrial lineage. The distinctive mitochondrial genome haplogroups, together with the shrinking and increasingly isolated occurrences of mammoths (Fig. 4), hint that Siberian mainland mammoths experienced a similar fate to those on Wrangel and St Paul Islands. However, whether the precise causes of their disappearance were the same[4,30], and whether the mainland mammoth also accumulated detrimental mutations consistent with genetic decline[31], will require further data to resolve.

The reliability of our method was further corroborated on the horse phylogeny (Supplementary Information 14.2). Successful assignment of ancient eDNA data to mitochondrial haplogroups, even when the

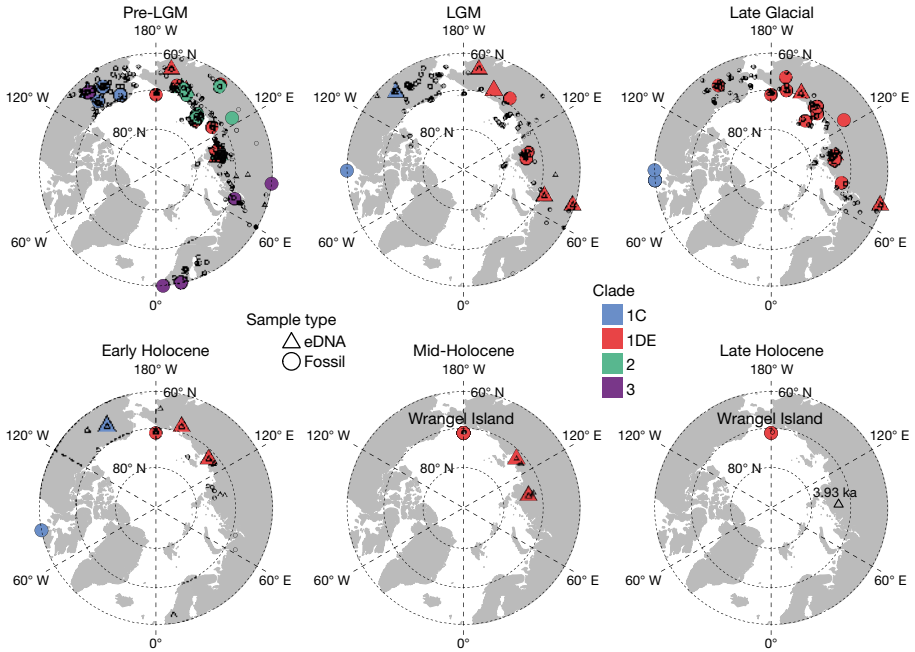


Fig. 4 | Mammoth distribution and mitochondrial haplotypes. A total of 78 mammoth mitochondrial genomes and 159 eDNA-identified mammoths (79 among them were assigned to mitochondrial haplotypes) are shown. Records of dated mammoth fossils[62] are also plotted. All samples older than 26.5 ka were combined into the pre-LGM interval.

DNA is highly degraded, highlights the potential for applying eDNA analysis to uncover population histories in regions in which fossils are rare or absent.

Concluding remarks

Controversy has persisted for decades over the nature of the Mammoth Steppe, a distinctive, now-vanished biome dominated by large mammal grazers[1,19,32]. Some studies, emphasizing the abundance of grazers (and the absence of large browsers), suggest that broad swaths of the unglaciated Late Pleistocene Arctic were covered by an extensive steppe dominated by low-sward herbaceous plants that were well-suited for megafaunal grazers[19,32]. Others, on the basis of pollen and plant macrofossil records, suggest that Arctic vegetation during this period was regionally diverse and included both tundra and steppe taxa[3,33]. Our results suggest the nature of the Mammoth Steppe lies in between these two seemingly conflicting interpretations. Consistent with the view of the Mammoth Steppe as a biome of intercontinental extent, our data show that various regions of the Arctic supported a more homogenous vegetation cover before and during the LGM (Extended Data Fig. 1b, c). We also found evidence of an elevated and episodic turnover of plant taxa during the Late Pleistocene compared with during the Holocene (Fig. 2a), consistent with inferences about changeable vegetation types during the glacial age based on the network of palaeobotanical (and fossil insect) sites presently available[3,12]. Jointly, our results suggest that the Mammoth Steppe was a regionally complex cryo-arid steppe, composed of forbs, graminoids and willow shrubs.

Our findings relating to the late survival of megafauna have important implications for the debate over the causes of Late Quaternary extinctions. Megafaunal survival into the Holocene indicates that, at least in certain parts of the Arctic and Subarctic, humans coexisted with these species for tens of thousands of years, which implies that human hunting was not an important factor in their extinction[6,25]. Instead, our results suggest that their extinction came when the last pockets of the steppe–tundra vegetation finally disappeared, when the Arctic-wide paludification was brought on by warmer and wetter climates[7,20].

What we have mined from this substantial dataset does not exploit its full potential. For example, we detected DNA of Camelidae (most probably the Arctic camel[34]) and *Panthera* (possibly the steppe lion). However, due to a lack of reference genomes for these species, we could not confirm these identifications. This constraint also applies to other species because our reference database—large as it is—is far from complete, despite our extensive sequencing efforts. With more species sequenced and new bioinformatics methods developed, this dataset can be reanalysed to explore more questions of Arctic biotic history.

Our study demonstrates how metagenomic analysis of eDNA extracted from ancient sediments can provide diverse insights, from detailed records of past flora and fauna to reconstructions of population histories and biotic interactions, to a greatly expanded spatiotemporal network of palaeoecological records. These advances are important in the context of continuous efforts to elucidate the past 50 kyr of Arctic biotic dynamics, especially given that the coevolution of plant and animal species, and their responses to the past climatic changes across this vast region, have previously been challenging to address at this resolution and at this scale using classical palaeobotanical and palaeontological data.

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/s41586-021-04016-x.

1. Binney, H. et al. Vegetation of Eurasia from the last glacial maximum to present: key biogeographic patterns. *Quat. Sci. Rev.* **157**, 80–97 (2017).
2. Clark, P. U. et al. The Last Glacial Maximum. *Science* **325**, 710–714 (2009).
3. Bigelow, N. H. Climate change and Arctic ecosystems: 1. Vegetation changes north of 55°N between the last glacial maximum, mid-Holocene, and present. *J. Geophys. Res.* **108**, https://doi.org/10.1029/2002jd002558 (2003).
4. Graham, R. W. et al. Timing and causes of mid-Holocene mammoth extinction on St. Paul Island, Alaska. *Proc. Natl Acad. Sci. USA* **113**, 9310–9314 (2016).

5. Stuart, A. J. Late Quaternary megafaunal extinctions on the continents: a short review. *Geol. J.* **50**, 338–363 (2015).
6. Koch, P. L. & Barnosky, A. D. Late Quaternary extinctions: state of the debate. *Ann. Rev. Ecol. Evol. Syst.* **37**, 215–250 (2006).
7. Rabanus-Wallace, M. T. et al. Megafaunal isotopes reveal role of increased moisture on rangeland during late Pleistocene extinctions. *Nat. Ecol. Evol.* **1**, 0125 (2017).
8. Mann, D. H., Groves, P., Kunz, M. L., Reanier, R. E. & Gaglioti, B. V. Ice-age megafauna in Arctic Alaska: extinction, invasion, survival. *Quat. Sci. Rev.* **70**, 91–108 (2013).
9. Capo, E. et al. Lake sedimentary DNA research on past terrestrial and aquatic biodiversity: overview and recommendations. *Quaternary* **4**, https://doi.org/10.3390/quat4010006 (2021).
10. Edwards, M. E. et al. Metabarcoding of modern soil DNA gives a highly local vegetation signal in Svalbard tundra. *Holocene* **28**, 2006–2016 (2018).
11. Hughes, P. D., Gibbard, P. L. & Ehlers, J. Timing of glaciation during the last glacial cycle: evaluating the concept of a global 'Last Glacial Maximum' (LGM). *Earth Sci. Rev.* **125**, 171–198 (2013).
12. Willerslev, E. et al. Fifty thousand years of Arctic vegetation and megafaunal diet. *Nature* **506**, 47–51 (2014).
13. Rasmussen, S. O. et al. A new Greenland ice core chronology for the last glacial termination. *J. Geophys. Res.* **111**, https://doi.org/10.1029/2005jd006079 (2006).
14. Mangerud, J. The discovery of the Younger Dryas, and comments on the current meaning and usage of the term. *Boreas* **50**, 1–5 (2020).
15. Bauska, T. K. et al. Carbon isotopes characterize rapid changes in atmospheric carbon dioxide during the last deglaciation. *Proc. Natl Acad. Sci. USA* **113**, 3465–3470 (2016).
16. Wesser, S. D. & Armbruster, W. S. Species distribution controls across a forest-steppe transition: a causal model and experimental test. *Ecol. Monogr.* **61**, 323–342 (1991).
17. Rijal, D. P. et al. Sedimentary ancient DNA shows terrestrial plant richness continuously increased over the Holocene in northern Fennoscandia. *Sci. Adv.* **7**, eabf9557 (2021).
18. Birks, H. H. Aquatic macrophyte vegetation development in Kråkenes Lake, western Norway, during the late-glacial and early-Holocene. *J. Paleolimnol.* **23**, 7–19 (2000).
19. Guthrie, R. D. Origin and causes of the mammoth steppe: a story of cloud cover, woolly mammal tooth pits, buckles, and inside-out Beringia. *Quat. Sci. Rev.* **20**, 549–574 (2001).
20. Mann, D. H., Peteet, D. M., Reanier, R. E. & Kunz, M. L. Responses of an Arctic landscape to Lateglacial and early Holocene climatic changes: the importance of moisture. *Quat. Sci. Rev.* **21**, 997–1021 (2002).
21. Ritchie, M. in *Competition and Coexistence* (eds Sommer, U. & Worm, B.) 109–131 (Springer, 2002).
22. Signor, P. W., Lipps, J. H., Silver, L. & Schultz, P. in *Geological Implications of Impacts of Large Asteroids and Comets on the Earth* vol. 190 (eds Silver, L. T. & Schultz, P. H.) 291–296 (1982).
23. Haile, J. et al. Ancient DNA reveals late survival of mammoth and horse in interior Alaska. *Proc. Natl Acad. Sci. USA* **106**, 22352–22357 (2009).
24. Librado, P. et al. Tracking the origins of Yakutian horses and the genetic basis for their fast adaptation to subarctic environments. *Proc. Natl Acad. Sci. USA* **112**, E6889–E6897 (2015).
25. Nikolskiy, P. A., Sulerzhitsky, L. D. & Pitulko, V. V. Last straw versus Blitzkrieg overkill: climate-driven changes in the Arctic Siberian mammoth population and the Late Pleistocene extinction problem. *Quat. Sci. Rev.* **30**, 2309–2328 (2011).
26. Pavlov, P., Svendsen, J. I. & Indrelid, S. Human presence in the European Arctic nearly 40,000 years ago. *Nature* **413**, 64–67 (2001).
27. Kuzmin, Y. V. & Keates, S. G. Siberia and neighboring regions in the Last Glacial Maximum: did people occupy northern Eurasia at that time? *Archaeol. Anthropol. Sci.* **10**, 111–124 (2016).
28. Stuart, A. J. & Lister, A. M. Extinction chronology of the woolly rhinoceros *Coelodonta antiquitatis* in the context of late Quaternary megafaunal extinctions in northern Eurasia. *Quat. Sci. Rev.* **51**, 1–17 (2012).
29. Chang, D. et al. The evolutionary and phylogeographic history of woolly mammoths: a comprehensive mitogenomic analysis. *Sci. Rep.* **7**, 44585 (2017).
30. Vartanyan, S. L., Arslanov, K. A., Karhu, J. A., Possnert, G. & Sulerzhitsky, L. D. Collection of radiocarbon dates on the mammoths (*Mammuthus primigenius*) and other genera of Wrangel Island, northeast Siberia, Russia. *Quat. Res.* **70**, 51–59 (2017).
31. Rogers, R. L. & Slatkin, M. Excess of genomic defects in a woolly mammoth on Wrangel island. *PLoS Genet.* **13**, e1006601 (2017).
32. Zimov, S. A., Zimov, N. S., Tikhonov, A. N. & Chapin, F. S. Mammoth steppe: a high-productivity phenomenon. *Quat. Sci. Rev.* **57**, 26–45 (2012).
33. Yurtsev, B. A. The Pleistocene "Tundra-Steppe" and the productivity paradox: the landscape approach. *Quat. Sci. Rev.* **20**, 165–174 (2001).
34. Rybczynski, N. et al. Mid-Pliocene warm-period deposits in the High Arctic yield insight into camel evolution. *Nat. Commun.* **4**, 1550 (2013).

Yucheng Wang[1,2,35], Mikkel Winther Pedersen[2,35], Inger Greve Alsos[3,35], Bianca De Sanctis[1,4], Fernando Racimo[2], Ana Prohaska[1], Eric Coissac[3,5], Hannah Lois Owens[6], Marie Kristine Føreid Merkel[3], Antonio Fernandez-Guerra[2], Alexandra Rouillard[2,7], Youri Lammers[3], Adriana Alberti[8,9], France Denoeud[9], Daniel Money[1], Anthony H. Ruter[2], Hugh McColl[2], Nicolaj Krog Larsen[2], Anna A. Cherezova[10,11], Mary E. Edwards[12,13], Grigory B. Fedorov[10,11], James Haile[2], Ludovic Orlando[14], Lasse Vinner[2], Thorfinn Sand Korneliussen[2,15], David W. Beilman[16], Anders A. Bjørk[17], Jialu Cao[2], Christoph Dockter[18], Julie Esdale[19], Galina Gusarova[3,20], Kristian K. Kjeldsen[21], Jan Mangerud[22,23], Jeffrey T. Rasic[24], Birgitte Skadhauge[18], John Inge Svendsen[22,23], Alexei Tikhonov[25], Patrick Wincker[9], Yingchun Xing[26], Yubin Zhang[27], Duane G. Froese[28], Carsten Rahbek[6,29], David Nogues Bravo[6], Philip B. Holden[30], Neil R. Edwards[30], Richard Durbin[4], David J. Meltzer[2,31], Kurt H. Kjær[2], Per Möller[32] & Eske Willerslev[1,2,33,34]✉

[1]Department of Zoology, University of Cambridge, Cambridge, UK. [2]Lundbeck Foundation GeoGenetics Centre, GLOBE Institute, University of Copenhagen, Copenhagen, Denmark. [3]The Arctic University Museum of Norway, UiT— The Arctic University of Norway, Tromsø, Norway. [4]Department of Genetics, University of Cambridge, Cambridge, UK. [5]Université Grenoble Alpes, Université Savoie Mont Blanc, CNRS, LECA, Grenoble, France. [6]Center for Macroecology, Evolution and Climate, GLOBE Institute, University of Copenhagen, Copenhagen, Denmark. [7]Department of Geosciences, UiT—The Arctic University of Norway, Tromsø, Norway. [8]Université Paris-Saclay, CEA, CNRS, Institute for Integrative Biology of the Cell (I2BC), Gif-sur-Yvette, France. [9]Génomique Métabolique, Genoscope, Institut François Jacob, CEA, CNRS, Université Evry, Université Paris-Saclay, Evry, France. [10]Institute of Earth Sciences, St Petersburg State University, St Petersburg, Russia. [11]Arctic and Antarctic Research Institute, St Petersburg, Russia. [12]School of Geography and Environmental Science, University of Southampton, Southampton, UK. [13]Alaska Quaternary Center, University of Alaska Fairbanks, Fairbanks, AK, USA. [14]Centre d'Anthropobiologie et de Génomique de Toulouse, Université Paul Sabatier, Faculté de Médecine Purpan, Toulouse, France. [15]National Research University, Higher School of Economics, Moscow, Russia. [16]Department of Geography and Environment, University of Hawaii, Honolulu, HI, USA. [17]Department of Geosciences and Natural Resource Management, University of Copenhagen, Copenhagen, Denmark. [18]Carlsberg Research Laboratory, Copenhagen, Denmark. [19]Center for Environmental Management of Military Lands, Colorado State University, Fort Collins, CO, USA. [20]Faculty of Biology, St Petersburg State University, St Petersburg, Russia. [21]Department of Glaciology and Climate, Geological Survey of Denmark and Greenland, Copenhagen, Denmark. [22]Department of Earth Science, University of Bergen, Bergen, Norway. [23]Bjerknes Centre for Climate Research, Bergen, Norway. [24]US National Park Service, Gates of the Arctic National Park and Preserve, Fairbanks, AK, USA. [25]Zoological Institute, , Russian Academy of Sciences, St Petersburg, Russia. [26]Resource and Environmental Research Center, Chinese Academy of Fishery Sciences, Beijing, China. [27]College of Plant Science, Jilin University, Changchun, China. [28]Department of Earth and Atmospheric Sciences, University of Alberta, Edmonton, Alberta, Canada. [29]Center for Global Mountain Biodiversity, GLOBE Institute, University of Copenhagen, Copenhagen, Denmark. [30]School of Environment, Earth and Ecosystem Sciences, The Open University, Milton Keynes, UK. [31]Department of Anthropology, Southern Methodist University, Dallas, TX, USA. [32]Department of Geology, Quaternary Sciences, Lund University, Lund, Sweden. [33]Wellcome Trust Sanger Institute, Wellcome Genome Campus, Cambridge, UK. [34]MARUM, University of Bremen, Bremen, Germany. [35]These authors contributed equally: Yucheng Wang, Mikkel Winther Pedersen, Inger Greve Alsos. ✉e-mail: ew482@cam.ac.uk

Methods

Sampling, chronology and eDNA taphonomy

Sampling and subsampling methods are described in Supplementary Information 1. Sample ages were determined through conventional or accelerator mass spectrometer radiocarbon (^{14}C) as well as optically stimulated luminescence. In total, 631 radiocarbon ages and 81 optically stimulated luminescence dates were used. For sedimentary sections with multiple contiguous dates without stratigraphic inversions, age–depth models were built to calculate sedimentation rates and estimate the ages of undated samples within these sections. All radiocarbon ages are in calibrated years before present, calibrated using IntCal20 (ref. [35]). Chronological information is provided in Supplementary Information 2 and Supplementary Data 1 and 2.

To determine whether DNA was in situ, control samples were obtained from modern surfaces, from water in adjacent rivers and lakes, and from stratigraphic layers bracketing the samples. Consistent with previous eDNA studies in the Arctic[12,23,36], we found no evidence of DNA leaching or redeposition in either terrestrial or lake sediment samples (Supplementary Information 5).

DNA extraction and sequencing

We tested the performance of different operations included in the widely used ancient eDNA extraction protocols[36–38] and a variety of purification methods on different sediment sample types. On the basis of these tests, we developed two new eDNA-extraction protocols that were optimized for isolating and purifying eDNA from our sediment samples (Supplementary Information 6.1 and 6.2). The InhibitEx-based protocol was then applied for extracting DNA from all samples. DNA extracts were thereafter converted into sequencing libraries according to the standard protocol[39], and sequenced using Illumina platforms after quality controls (Supplementary Information 6.3). All DNA extractions and pre-index analyses were performed in the dedicated ancient DNA laboratories at the Centre for GeoGenetics, University of Copenhagen, according to established ancient DNA protocols[40].

PhyloNorway plant genome database construction

The PhyloNorway plant genome database was constructed by sequencing 1,541 Arctic and boreal plant specimens collected from herbaria. DNA was extracted from the selected specimens using a modified Macherey–Nagel Nucleospin 96 Plant II protocol. Two different library preparation protocols were applied depending on DNA yields. All of the libraries were then sequenced. Nuclear ribosomal DNA and chloroplast genome from each plant were assembled to evaluate the data quality. Whole-genome contigs for each plant were assembled and annotated as the final reference database. A list of plant species, herbarium information, DNA extraction, sequencing and database statistics are supplied in Supplementary Data 3. Data for three standard barcodes skimmed from this database were also used in ref. [41]. Details are provided in Supplementary Information 7.

Taxonomic identification, authentication and quantification

We performed taxonomic classification by mapping reads against a comprehensive genomic database that was annotated with taxonomic information according to the principle of the Holi pipeline[36]. Details of the composition of the reference database are provided in Supplementary Information 9.2.1.

All reads were first quality-controlled, and each read was then offered an equal chance to be aligned against all entries in the database after duplicate removal (Supplementary Information 9.1 and 9.2). No limitation to specific taxonomic group, geography or environment was applied for the alignment. The lowest common ancestor of all of the hits with 100% similarity was assigned to each read that had been aligned to multiple taxa. The taxonomic coverage of different database compositions and their effects on taxa identification were evaluated using a *k*-mer-based method (Supplementary Information 9.2.2). We found that using a proper reference database is important for eDNA metagenomics-based taxa identification, particularly for ancient datasets in which the DNA is highly fragmented. Even reference genome availability across taxa can improve the sensitivity and specificity of the identification by increasing the identified reads and correcting the misidentifications (Supplementary Information 9.2.5). Taxa that were detected in the laboratory controls were combined into a list, and all of the listed taxa were subtracted from samples (Supplementary Information 9.3). The resulting plant and animal taxonomic profiles were thereafter parsed for additional authentication using a series of conservative thresholds (Supplementary Information 9.4 and 9.6), on the basis of an Arctic flora and faunal checklist (Supplementary Information 8). Plant taxa that passed these filters all have Arctic or boreal distributions (Supplementary Information 9.4). All eDNA reads aligned to an animal were further confirmed as exclusive alignments, by requiring perfect alignment to that animal, and no alignment to any other organisms when allowing for 1 or 2 mismatches (Supplementary Information 9.6.3). The two extinct animals— mammoth and woolly rhinoceros— were also confirmed by the DNA-damage patterns (Supplementary Information 9.6.2).

Relative abundances for plants were estimated on the basis of the number of the assigned reads, by excluding the effects of DNA degradation in different samples, and eliminating the effects of the sequencing depth among different samples and the efficiency of the taxa-identification pipeline among different taxa (Supplementary Information 9.5).

Vegetation diversity and dissimilarity

The Shannon diversity index was calculated according to the method in ref. [42]. Plant morphological forms were assigned at the genus level on the basis of the plant trait database of eFloras (http://www.efloras.org). Beta-diversity (dissimilarity) between every two plant assemblages was calculated according to the method in ref. [43]. For the pan-Arctic vegetation turnover (Fig. 2a), plant genera identified in all samples in each 2,000-year interval were combined as an assemblage; beta-diversity between each two consecutive intervals was calculated. Regional vegetation turnover (Fig. 2c) was calculated at 5,000-year intervals. NMDS ($k = 3$; Extended Data Fig. 1c) was performed using the R package vegan[44], allowing 100,000 iterations of random starting to find the best convergent solution. Correlations between the abundance of each plant genus (or proportion of each morphological form) and the values of each of the three NMDS components (Extended Data Fig. 4b) were assessed using the Pearson product–moment correlation and *t*-test ($P < 0.05$).

Comparison of eDNA shotgun metagenomics and metabarcoding

We applied two modules for comparing the metabarcoding and shotgun metagenomics in taxa identifications. (1) We conducted the two sequencing techniques in parallel on 14 DNA extracts to directly compare the retrieved taxonomic profiles. (2) We compared the floristic profiles reconstructed by this study and a previous metabarcoding study[12] on 131 overlapping samples of the two datasets. The results show that metagenomics performed better on our samples in both captured floristic and faunal diversity. Details are provided in Supplementary Information 11.

Palaeoclimate panels and human distribution niche modelling

For the ice-core data from Greenland (Fig. 2a), we rescaled the available δ^{18}O ratios (20-year slices) retrieved from NGRIP1 (ref. [45]), NGRIP2 (ref. [46]) and GISP2 (ref. [47]) to the range of the corresponding ratio of GRIP[48], for which there are valid values for all age slices, using the rescale function in the R package scales. The mean of the available ratios for each time slice from the four datasets was calculated and used. Calcium concentrations were calculated from refs. [49,50] using the same method as for δ^{18}O. Snow-accumulation rates were based on GISP2 (ref. [51]).

We also modelled monthly palaeoclimate anomalies at 1,000-year time steps using an emulator[52] and downscaled them onto a modern baseline climatology (CHELSA)[53] at a spatial resolution of 1°. From these data, we calculated four environmental variables—annual mean temperature, temperature seasonality, annual precipitation and precipitation seasonality—that were used to represent the climate for each of our eDNA sites. Details are provided in Supplementary Information 12.1.

We developed distribution models to map environmentally suitable conditions for Palaeolithic human occurrence in steps of 1,000 years from 5 ka to 31 ka and steps of 2,000 years from 32 ka to 47 ka. First, geo-references for human remains in the Arctic were collected and dates from ^{14}C calibrations inferred from two databases CARD2.0 (ref. [54]) and the Palaeolithic of Europe[55]. These data were filtered for quality, resulting in a final set of 6,497 occurrences. From 32 ka to 47 ka, we calculated 2,000-year averages of the four environmental variables. We then generated five-algorithm ensemble models at each time step to characterize the climatic niche of Palaeolithic humans. We validated all of the models by assessing the area under the receiver operating characteristic curve (AUC) and true skill statistic; we also used model AUCs to generate weighted ensemble models at each time step. Finally, we projected the ensemble models into geographic space to map climatic suitability for humans, expressed as the potential presence or absence at each time step at each of the eDNA sites. Details are provided in Supplementary Information 12.2.

Spatiotemporal models for animal eDNA

We combined our animal eDNA data with the modelled climate variables, projected human occurrence and the NMDS ordinations of vegetation to examine the relative impacts of climate, human activity and vegetation on the geographical distributions of a selected group of Arctic mammals. We developed a method to spatiotemporally model animal eDNA presence, using these three sets of variables, while accounting for auto-correlation in time and space. The method uses a hierarchical Bayesian model that includes a spatiotemporal Gaussian random field, and was implemented in R-INLA[56,57]. We used the Watanabe–Akaike information criterion to assess the model fit using different sets of covariates. Detailed methods are provided in Supplementary Information 13.

Mammoth and horse mitochondrial haplotyping

We placed eDNA mitochondrial reads for mammoth and horse into their respective mitochondrial reference phylogenies using recently developed software[58]. We used existing variation to assign informative markers onto branches of a mitochondrial phylogeny, then determined the number of supporting and conflicting single-nucleotide polymorphisms for each eDNA sample on each branch of the tree to place the sample onto the most likely branch. Detailed methods are provided in Supplementary Information 14.

Statistics and data visualization

Changing trends are illustrated against time (Fig. 2a–c and Extended Data Fig. 2b, c) or distance (Extended Data Fig. 1b) via the Loess Smooth (span = 4) function in the R package ggplot2 (ref. [59]), with original data points or confidence intervals (s.e.) shown when other curves are not obstructed. The heat maps showing the mean of a genus' proportions across all samples within an age interval were generated using the R package ComplexHeatmap[60]. The mammoth phylogenetic tree was illustrated using ggtree, which is included in the R package ggplot2. The base map source for Fig. 1 and Extended Data Fig. 5 was Arctic SDI and, for Fig. 4, was the R package maptools.

Reporting summary

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

Data availability

Adapter-removed plant or animal eDNA data were deposited at EMBL-ENA under project accession ERP127790. The raw data of PhyloNorway plant genome database are available at EMBL-ENA under project accession PRJEB43865. Assembled plant genome contigs of the PhyloNorway database are available at DataverseNO[61]. NCBI databases are available at the NCBI ftp server (https://ftp.ncbi.nlm.nih.gov). The Canadian Archaeological Radiocarbon Database (CARD2.0) is available online (https://www.canadianarchaeology.ca). The Radiocarbon Palaeolithic Europe Database is available online (https://ees.kuleuven.be/geography/projects/14c-palaeolithic). All other data are provided in the Supplementary Information and Supplementary Data 1–9.

Code availability

Scripts are archived at GitHub (https://github.com/wyc661217/Arctic_eDNA_2021).

35. Reimer, P. J. et al. The IntCal20 Northern Hemisphere radiocarbon age calibration curve (0–55 cal kBP). *Radiocarbon* **62**, 725–757 (2020).
36. Pedersen, M. W. et al. Postglacial viability and colonization in North America's ice-free corridor. *Nature* **537**, 45–49 (2016).
37. Slon, V. et al. Neandertal and Denisovan DNA from Pleistocene sediments. *Science* **356**, 605–608 (2017).
38. Lorenz, M. G. & Wackernagel, W. Adsorption of DNA to sand and variable degradation rates of adsorbed DNA. *Appl. Environ. Microb.* **53**, 2948–2952 (1987).
39. Meyer, M. & Kircher, M. Illumina sequencing library preparation for highly multiplexed target capture and sequencing. *Cold Spring Harb. Protoc.* **2010**, pdb.prot5448 (2010).
40. Willerslev, E., Hansen, A. J. & Poinar, H. N. Isolation of nucleic acids and cultures from fossil ice and permafrost. *Trends Ecol. Evol.* **19**, 141–147 (2004).
41. Alsos, I. G. et al. The treasure vault can be opened: large-scale genome skimming works well using herbarium and silica gel dried material. *Plants* **9**, https://doi.org/10.3390/plants9040432 (2020).
42. Hill, M. O. Diversity and evenness: a unifying notation and its consequences. *Ecology* **54**, 427–432 (1973).
43. Koleff, P., Gaston, K. J. & Lennon, J. J. Measuring beta diversity for presence-absence data. *J. Anim. Ecol.* **72**, 367–382 (2003).
44. Dixon, P. VEGAN, a package of R functions for community ecology. *J. Veg. Sci.* **14**, 927–930 (2003).
45. Grootes, P. M. & Stuiver, M. Oxygen 18/16 variability in Greenland snow and ice with 10^{-3}- to 10^{5}-year time resolution. *J. Geophys. Res. Oceans* **102**, 26455–26470 (1997).
46. Andersen, K. K. et al. High-resolution record of Northern Hemisphere climate extending into the last interglacial period. *Nature* **431**, 147–151 (2004).
47. Stuiver, M. & Grootes, P. M. GISP2 oxygen isotope ratios. *Quat. Res.* **53**, 277–284 (2017).
48. Johnsen, S. J. et al. The δ^{18}O record along the Greenland Ice Core Project deep ice core and the problem of possible Eemian climatic instability. *J. Geophys. Res. Oceans* **102**, 26397–26410 (1997).
49. Fuhrer, K., Neftel, A., Anklin, M. & Maggi, V. Continuous measurements of hydrogen peroxide, formaldehyde, calcium and ammonium concentrations along the new grip ice core from summit, Central Greenland. *Atmos. Environ. A* **27**, 1873–1880 (1993).
50. Mayewski, P. A. et al. Major features and forcing of high-latitude northern hemisphere atmospheric circulation using a 110,000-year-long glaciochemical series. *J. Geophys. Res. Oceans* **102**, 26345–26366 (1997).
51. Alley, R. B. et al. Abrupt increase in Greenland snow accumulation at the end of the Younger Dryas event. *Nature* **362**, 527–529 (1993).
52. Holden, P. B. et al. PALEO-PGEM v1.0: a statistical emulator of Pliocene–Pleistocene climate. *Geosci. Model Dev.* **12**, 5137–5155 (2019).
53. Karger, D. N. et al. Climatologies at high resolution for the earth's land surface areas. *Sci. Data* **4**, 170122 (2017).
54. Martindale, A. et al. Canadian Archaeological Radiocarbon Database (CARD 2.1) (Laboratory of Archaeology at the University of British Columbia, and the Canadian Museum of History, accessed 6 February 2020).
55. Vermeersch, P. M. Radiocarbon Palaeolithic Europe database: a regularly updated dataset of the radiometric data regarding the Palaeolithic of Europe, Siberia included. *Data Brief* **31**, 105793 (2020).
56. Rue, H., Martino, S. & Chopin, N. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *J. R. Stat. Soc. B* **71**, 319–392 (2009).
57. Lindgren, F. & Rue, H. Bayesian spatial modelling with R-INLA. *J. Stat. Softw.* **63**, 1–25 (2015).
58. Martiniano, R., De Sanctis, B., Hallast, P. & Durbin, R. Placing ancient DNA sequences into reference phylogenies. Preprint at https://doi.org/10.1101/2020.12.19.423614 (2020).
59. Wickham, H. *ggplot2: Elegant Graphics for Data Analysis* (Springer, 2016).
60. Gu, Z., Eils, R. & Schlesner, M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics* **32**, 2847–2849 (2016).
61. Wang, Y. et al. Supporting data for: Late Quaternary dynamics of Arctic Biota revealed by ancient environmental metagenomics. https://doi.org/10.18710/3CVQAG, DataverseNO, V1 (2021).
62. Theodoridis, S. et al. Climate and genetic diversity change in mammals during the Late Quaternary. Preprint at https://doi.org/10.1101/2021.03.05.433883 (2021).

Acknowledgements We thank D. H. Mann for his detailed and constructive comments; and T. Ager, J. Austin, T. B. Brand, A. Cooper, S. Funder, M. T. P. Gilbert, T. Jørgensen, N. J. Korsgaard, S. Liu, M. Meldgaard, P. V. S. Olsen, M. L. Siggaard-Andersen, J. Stenderup, S. A. Woodroffe and staff at the GeoGenetics Sequencing Core and National Park Service-Western Arctic National Parklands for help and support. E.W. and D.J.M. thank the staff at St. John's College, Cambridge, for providing a stimulating environment for scientific discussion of the project. E.W. thanks Illumina for collaboration. The Lundbeck Foundation GeoGenetics Centre is supported by the Carlsberg Foundation (CF18-0024), the Lundbeck Foundation (R302-2018-2155), the Novo Nordisk Foundation (NNF18SA0035006), the Wellcome Trust (UNS69906) and GRF EXC CRS Chair (44113220)—Cluster of Excellence. The PhyloNorway plant genome database is part of the Norwegian Barcode of Life Network (https://www.norbol.org) funded by the Research Council of Norway (226134/F50), the Norwegian Biodiversity Information Centre (14-14, 70184209) and The Arctic University Museum of Norway. Metabarcoding sequencing was funded by the Central Public-Interest Scientific Institution Basal Research Fund, CAFS (2017B001 and 2020A001). B.D.S. is supported by the Wellcome Trust programme in Mathematical Genomics and Medicine (WT220023); F.R. by a Villum Fonden Young Investigator award (no. 00025300); D.J.M. by the Quest Archaeological Research Fund; P.M. by the Swedish Research Council (VR); R.D. by the Wellcome Trust (WT207492); and A.R. by a Marie Skłodowska-Curie Actions Individual Fellowship (MSCA-IF, 703542) and the Research Council of Norway (KLIMAFORSK, 294929). L.O. has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (no. 681605); I.G.A. and Y.L. from the ERC under the European Union's Horizon 2020 research and innovation programme (no. 819192). J.I.S. and J.M. are supported by the Research Council of Norway. P.B.H. and N.R.E. acknowledge NERC funding (grant NE/P015093/1). D.W.B. was supported by a Marie Skłodowska-Curie Actions Incoming International Fellowship (MCIIF-40974). T.S.K. is funded by a Carlsberg Foundation Young Researcher Fellowship (CF19-0712).

Author contributions B.D.S., F.R., A.P. and E.C. contributed equally to this work. E.W. and Y.W. initiated and led the study. E.W., Y.W., M.W.P. and K.H.K. designed the study. P.M., K.H.K., E.W., I.G.A., Y.W., J.I.S., D.G.F., J.M., A.R., M.E.E., J.H., D.W.B., A.A.B., J.E., K.K.K., J.T.R., A.T., A.A.C. and G.B.F. provided samples. Y.W., N.K.L. and A.R. revised the age of samples. Y.W. performed the DNA laboratory work. Y.W., B.D.S., A.F.-G., R.D., D.M. and T.S.K. did the bioinformatics and statistical analyses. I.G.A., Y.W. and A.R. evaluated the plant taxonomic profiles. For the PhyloNorway plant genome database: I.G.A. and E.C. designed the study and led the work; M.K.F.M. and A.A. did the laboratory work; P.W. led the sequencing; E.C., Y.W., Y.L. and F.D. did the bioinformatics analyses. H.L.O., D.N.B., Y.W., C.R. and D.J.M. modelled the human distribution. P.B.H. and N.R.E. modelled the climate. F.R. developed the animal spatiotemporal model. Y.W., M.W.P., D.J.M., I.G.A., E.W. and A.P. drafted the manuscript which all of the co-authors commented on.

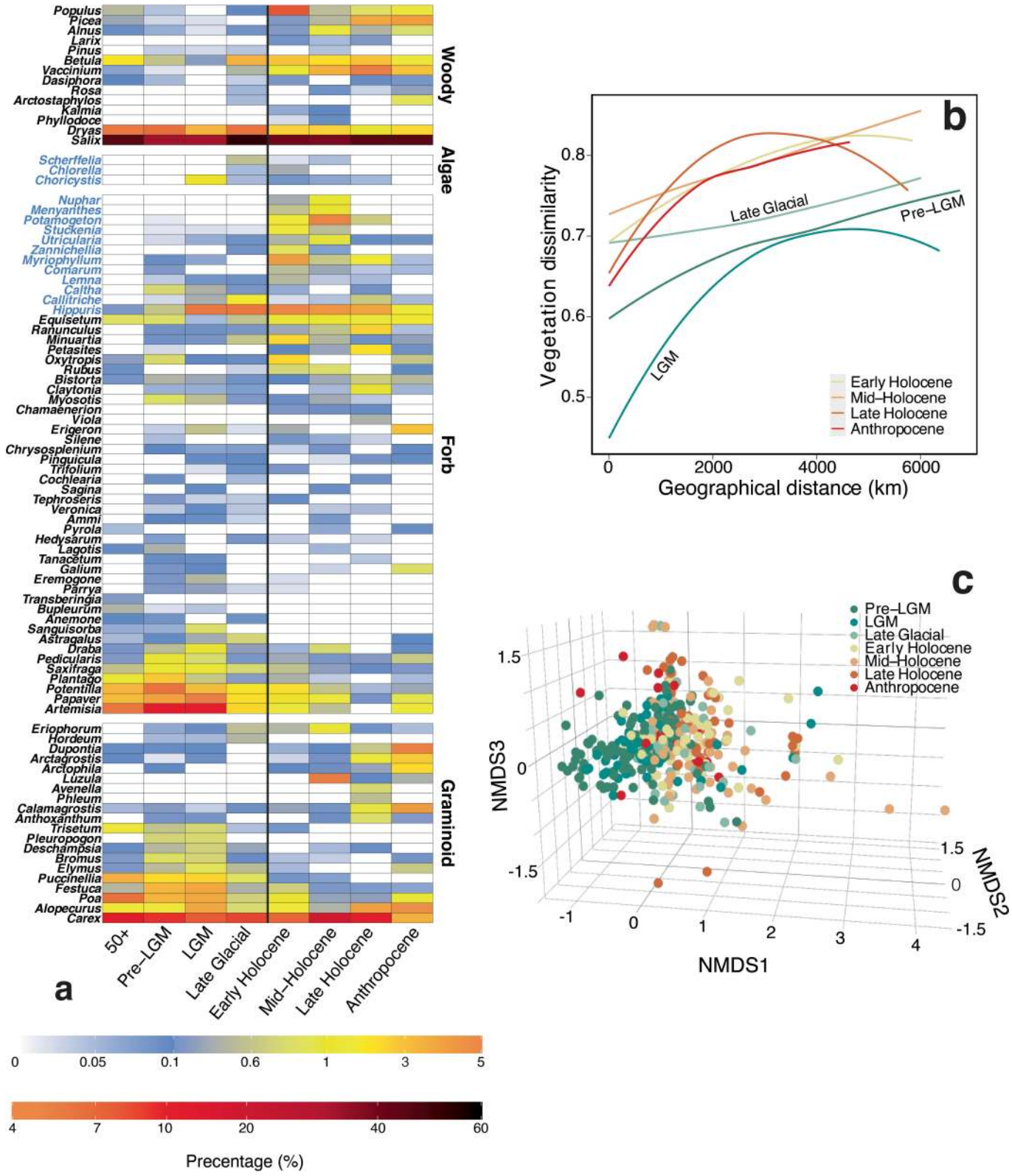
Competing interests The authors declare no competing interests.

Additional information

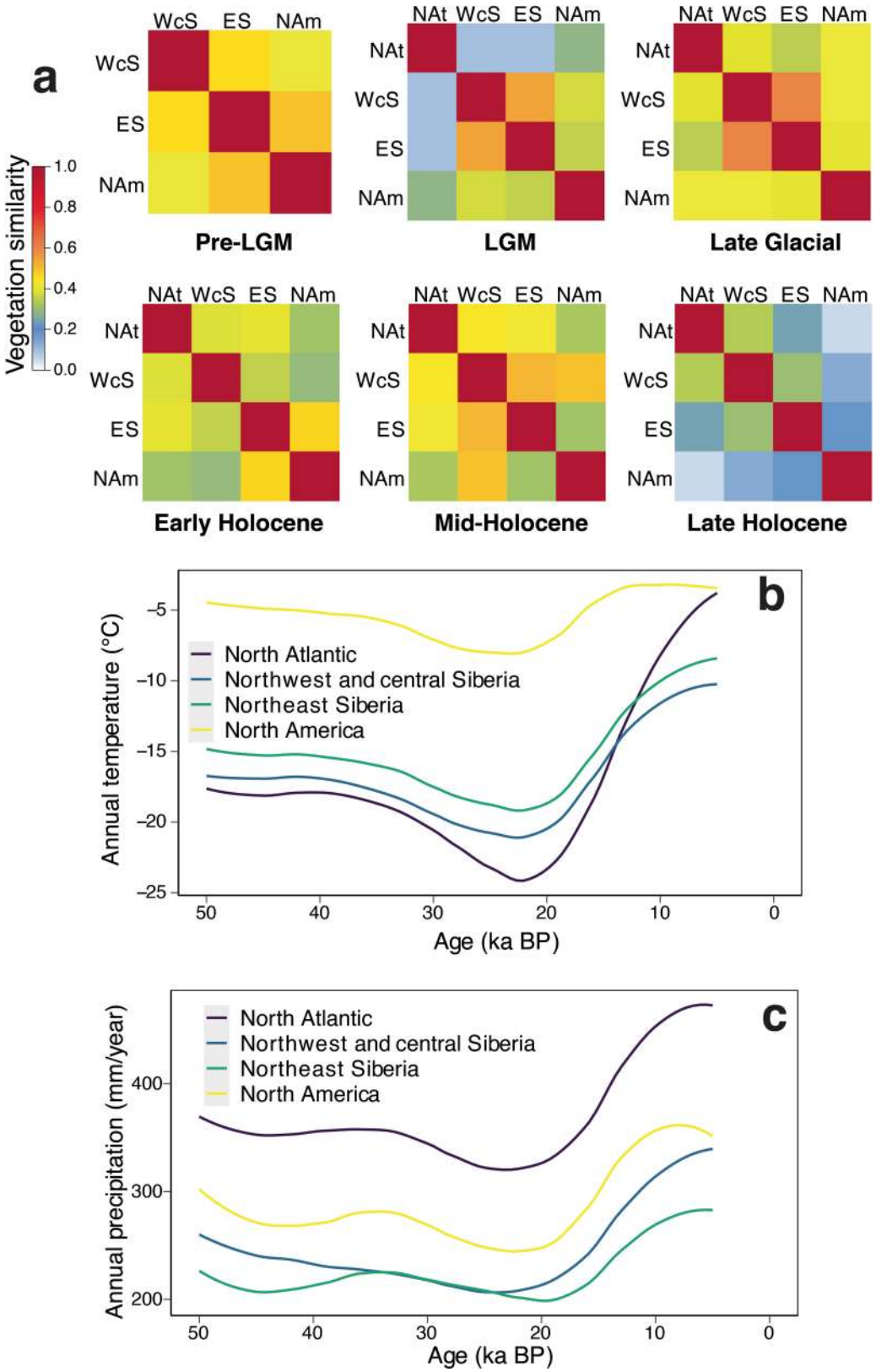
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41586-021-04016-x.

Correspondence and requests for materials should be addressed to Eske Willerslev.

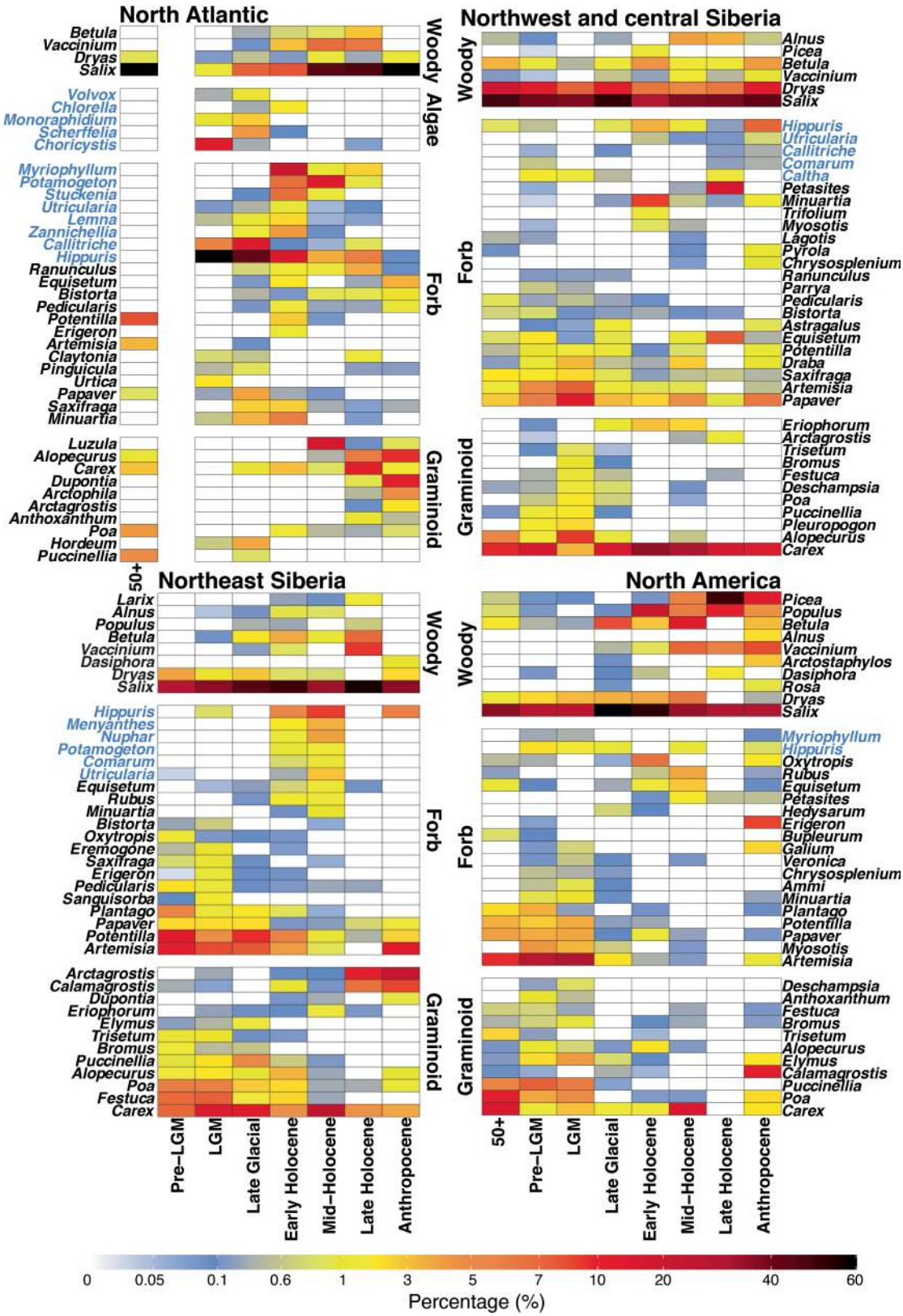
Peer review information *Nature* thanks Patricia Fall, Brian Huntley, Paul Valdes and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.



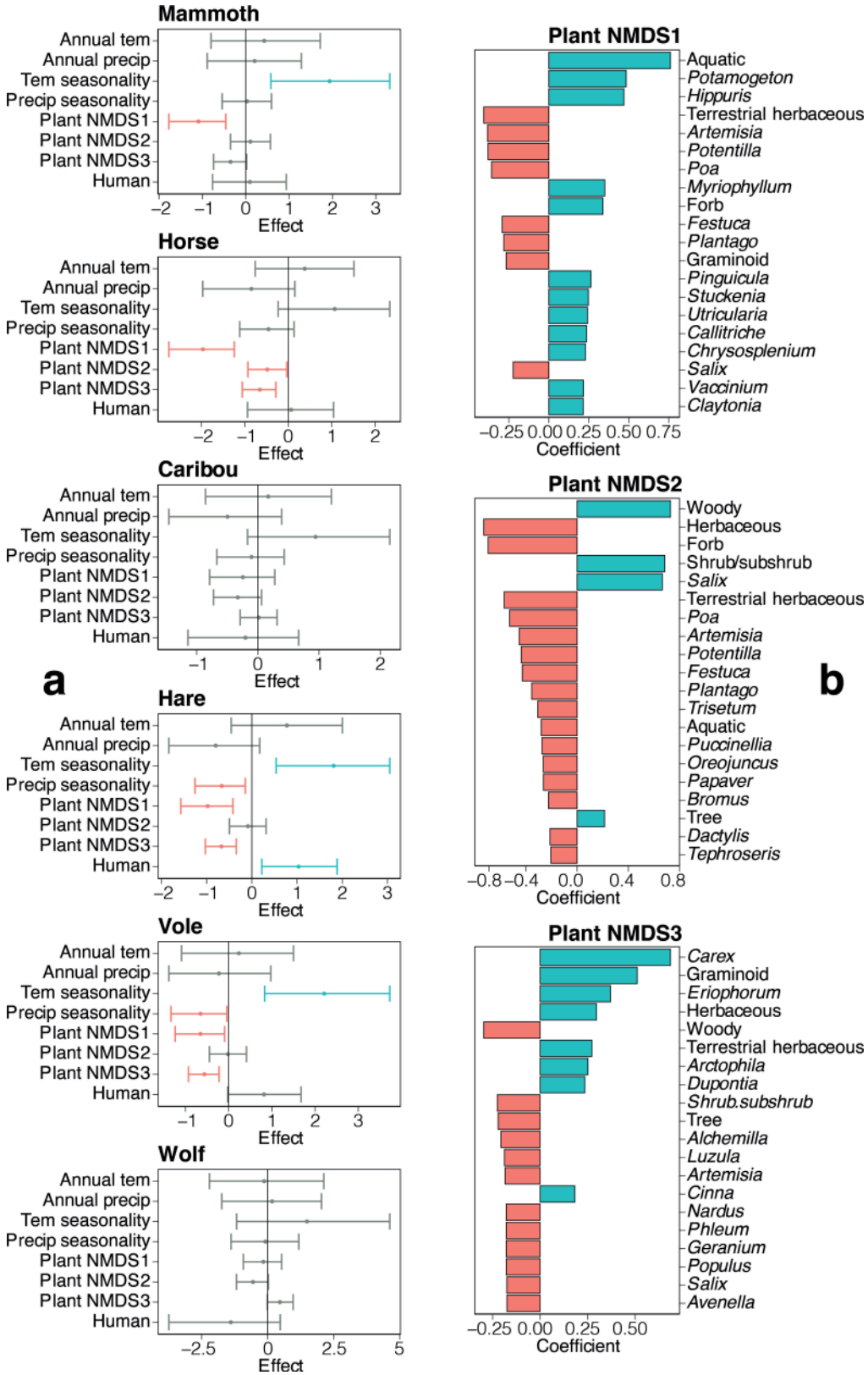
Extended Data Fig. 1 | Circum-Arctic plant abundance variations and vegetation similarity clustering. a, Pan-Arctic plant abundance heatmap. **b**, Spatial vegetation dissimilarities. Pairwise spatial beta-diversities (dissimilarities between every two plant communities) against the geographical distances between the two communities. **c**, Non-metric Multidimensional Scaling (NMDS, k=3) on vegetation communities.



Extended Data Fig. 2 | Regional vegetation differences and climate changes. a, Vegetation similarities between each two regions. All identified plant genera across sites in a region during a time interval were merged as a plant assemblage. Spatial beta-diversity between every two assemblages were calculated and illustrated. NAt, North Atlantic; WcS, Northwest and central Siberia; ES, Northeast Siberia; Nam, North America. **b** and **c**, Modelled annual temperature and precipitation in different regions. Means of the modelled annual temperature and precipitation values (Methods) at all eDNA sampling sites within a region at each 1,000-year time step were calculated. The changing trends are illustrated.

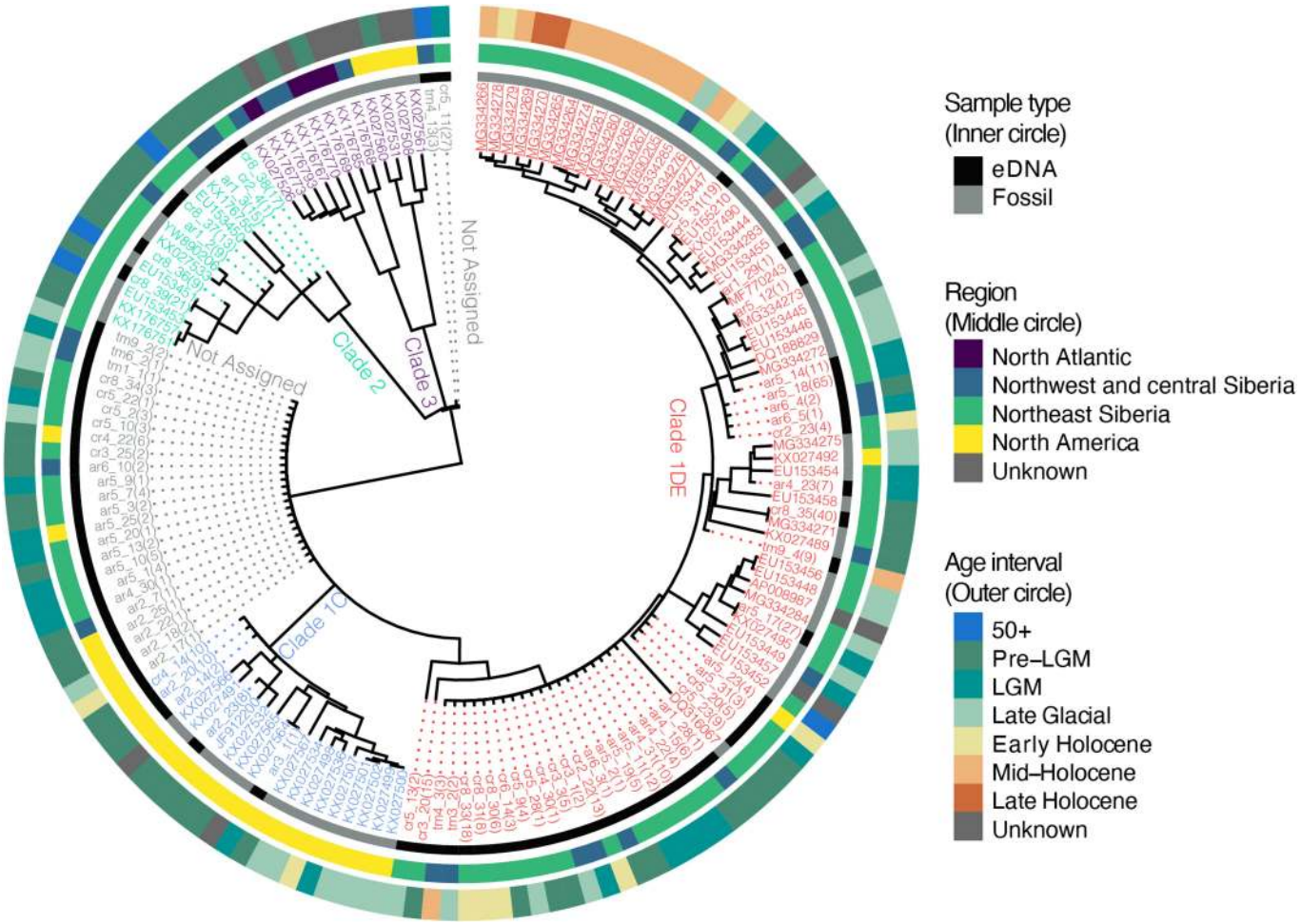


Extended Data Fig. 3 | Regional plant abundance heatmaps. Heatmaps show the relative abundances of the 40 abundant plant genera in each region.



Extended Data Fig. 4 | Environmental explanatory factors for animal distribution, and plant NMDS components. a, Posterior parameter estimates of covariate effects for the models explaining the presence/absence of each animal's eDNA using climate, human presence and plant NMDS as explanatory variables. The dots represent the posterior means, and the whiskers represent the posterior 2.5% and 97.5% quantiles. The colour red denotes covariate effects whose 2.5% and 97.5% quantiles are both negative, while the colour blue denotes covariate effects 2.5% and 97.5% quantiles are both positive. **b**, The plant genera and morphological forms correlated to the 3 components of plant NMDS. Plant genera (morphological forms) are ranked by the p-value of t-test, and only the top 20 Pearson correlations are shown. The colour red denotes negative correlations while the colour blue denotes positive correlations.

Extended Data Fig. 5 | Distribution chronologies for woolly rhinoceros, bison, horse, caribou, hare, wolf, and vole. We combined our DNA results and the fossil records[62] (available for woolly rhinoceros, bison, and caribou). Samples older than 26.5 ka were combined into Pre-LGM; samples younger than 4.2 ka were combined into the Late Holocene.



Extended Data Fig. 6 | Mammoth mitochondrial phylogenetic tree. For placed eDNA samples the number of supporting single-nucleotide polymorphisms is given in braces (Methods). IDs for the Wrangel Island population are underlined.

nature research

Corresponding author(s): Michaela Frye

Last updated by author(s): May 5, 2022

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☐	☒	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	IVIS Spectrum bioluminescence images were acquired with Living Image v4.4 (PerkinElmer). Confocal microscope images were acquired with LAS X v1.8.1.13759 (Leica). Flow cytometry data were collected using BD FACSDiva 8.0. Electron Microscopy of cell sections were observed in a Zeiss EM 910 at 120kV (Carl Zeiss, Oberkochen, Germany) and micrographs taken using a slow scan CCD camera (TRS, Moorenweis, Germany). The collection of the oxygen consumption rate and acidification rate values were done on the software Wave Desktop 2.6.1. For RT-qPCR, QuantStudio 5 qPCR software (Applied Biosystems) was used. Metabolic data was acquired using Empower3 software suite (Waters).
Data analysis	Graphpad Prism 9 was used for the statistics and generation of graphs. Flow cytometry data analysis was done with FlowJo V10 (Treestar). Bioluminescence data were analyzed using Living Image software V4.4. Finch TV1.5 was used to analyse bisulfite treated sequences. Proteomics raw files were analyzed using Proteome Discoverer (PD) 2.4 software (ThermoFisher Scientific), SequestHT node and Percolator. Gene Ontologies were determined using GSEA v4.0.3 software, ToppGene (https://toppgene.cchmc.org/), or GOrilla (http://cbl-gorilla.cs.technion.ac.il/). Clusteranalyses was done using Clustvis 2.0 (https://biit.cs.ut.ee/clustvis/). Quantification of signal from immunofluorescence images was done with ImageJ2 FIJI Version 2.1.0/153c. RNAseq analysis used the following software: FastX toolkit 0.0.13; Homertools 4.7; STAR v2.3; PicardTools 1.78; featureCounts v1.4.5-p1; R v3.3.2 EdgeR v3.16; DESeq2(v1.4.1). For in silico analyses of TCGA-HNSCC tumours, the following softwares were used: EdgeR, "GSVA" package in R, maxstat in R. For Bisulfite sequencing analysis, the sequences were trimmed with 'TrimGalore !' and aligned with Bismark. The R packages RSamtools, GenomicAlignments, and VIM.

Custom perl script to compare groups in RNA-seq input table can be found here: https://zenodo.org/search?page=1&size=20&q=6518420

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Quantitative proteomics data are available on PRIDE (PXD021835). RNA sequencing data using VDH01 and VDH15 cells are available on EGA under the accession number EGAS00001004765 including the attached studies EGAD00001008743 and EGAD00001008742. All other sequencing data are deposited on GEO under the accession number GSE201993. Results are in part based on TCGA-HNSC (accession number phs000178) downloaded from TCGA (https://portal.gdc.cancer.gov).

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	The number of orthotopic transplantation assays per condition was chosen to be a minimum of 5 biological replicates per condition. This number is based on previous experiences in animal models and publications (e.g. Pascual et al., Nature, 2017). All experiments were generally powered to detect differences greater than 20% at a significance of $p<0.05$. For targeted and genome-wide gene expression analyses of cultured cells, our study used a minimum of 3 replicates (one replicate = one independent infection). This sample size was estimated using previous NGS datasets (e.g. Blanco et al. 2016 Nature; Selmi et al. 2021 NAR). Flow cytometry, RT-qPCRs and seahorse assays were additionally performed in at least 3 technical replicates.
Data exclusions	No data were excluded.
Replication	All experiments were repeated at least twice and all repeats were successful.
Randomization	Experimental animals were randomly assigned to each experimental cohorts. For cell culture experiments, cells were equally distributed into multi-plate wells and the treatment condition was randomly applied. No additional controls for covariates was performed as the mice were age- and gender-matched and cells from the same passages were used for the experiments.
Blinding	Animal, cellular, flow cytometry and fluorescent quantifications were performed in a blinded manner. Bisulfite sequenced samples were processed in a blinded manner.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☐	☒ Antibodies
☐	☒ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☐	☒ Animals and other organisms
☐	☒ Human research participants
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☐	☒ Flow cytometry
☒	☐ MRI-based neuroimaging

Antibodies

Antibodies used: CD44 (Thermofisher, #14-0441-82, IM7, Lot#2093235 , 1:200), MTCO2 (Abcam, ab91317, Rabbit polyclonal, 1:200), MTCO1

Antibodies used	(Thermofisher, Rabbit Polyclonal, PA5-26688, Lot#SA100601BX, 1:200), cytokeratin 10 (Biolegend, PRB-159P, Poly19054, Lot#B284664, 1:200) and filaggrin (Covance, PRB-417P-100, Poly19058, Lot#B257576, 1:200), HSP90 (Santa Cruz, sc-13119, F-8, Lot#J2616, 1:1000) . NSUN3 (Genetex, GTX46175, Rabbit Polyclonal, Lot#822004446 1:100) or anti-GLUT1 (Abcam, Ab15309, Rabbit polyclonal, Lot#GR3266142-1, 1:100). The specificity of antibody staining was confirmed by IHC staining with using a rabbit IgG isotype control antibody (Cell Signaling Technology, #3900, DA1E, 1:1000). PE-Cy7-conjugated CD44 (BD Pharmingen, #560533, G44-26, Lot#0037983, 1:300), FITC- or eFluor 660- conjugated CD36 (BD Bioscience, 1:500, , #555454 and #50-0369-42, NL07, Lot#2303260, Thermofisher).
Validation	All antibodies are commercially available and have been validated in previously published studies: Anti-CD44 (#14-0441-82, Thermofisher). This monoclonal antibody recognizes all human CD44 isoforms (clone IM7) and is reported in 213 studies. https://www.thermofisher.com/antibody/product/CD44-Antibody-clone-IM7-Monoclonal/14-0441-82 anti-MTCO2 (ab91317, Abcam). This polyclonal antibody recognizes the human MT-CO2, is validated by the manufacturer for WB application and is used in 7 different pubblished studies. https://www.abcam.com/mtco2-antibody-ab91317.html anti-MTCO1 (PA5-26688, Thermofisher). This polyclonal antibody recognizes the human MTCO1 protein, is validated by the manufacturer for WB, IHC and immunofluorescence applications. https://www.thermofisher.com/antibody/product/MTCO1-Antibody-Polyclonal/PA5-26688 Anti-Cytokeratin 10 (PRB-159P, Bioilegend). This polyclonal antibody recognizes the human protein cytokeratin 10 and is validated by the manufacturer. https://www.biolegend.com/en-us/products/keratin-10-polyclonal-antibody-purified-10952 Anti-filaggrin (PRB-417P-100, Covance). This polyclonal antibody recognizes the human protein of filaggrin and is validated by the manufacturer. https://www.biolegend.com/en-us/products/filaggrin-polyclonal-antibody-purified-10943 anti-NSUN3 (Genetex, GTX46175). This polyclonal antibody recognizes the human protein NSUN3 and has been validated by us through NSUN3-knock-down experiments in this study and the manufacturer for WB application. https://www.genetex.com/Product/Detail/NSUN3-antibody-C-term/GTX46175 anti-GLUT1 (Abcam, Ab15309). This polyclonal antibody recognizes the human GLUT1 protein and has been used in 103 different published studies. https://www.abcam.com/glucose-transporter-glut1-antibody-ab15309.html anti-HSP90 ((Santa Cruz, sc-13119). This monoclonal antibody recognizes the human HSP90 protein and has been used in 567 different published studies. https://www.scbt.com/fr/p/hsp-90alpha-beta-antibody-f-8 IgG isotype control antibody (DA1E, Cell Signaling Technology). This isotype control antibody are used to estimate the non specific binding of target primary antibodies due to Fc receptor binding or protein-protein interaction. It is validated by the manufacturer. https://www.cellsignal.com/products/primary-antibodies/rabbit-da1e-mab-igg-xp-isotype-control/3900?Ntk=Products&Ntt=3900 PE-Cy7-conjugated CD44 (BD Pharmingen, #560533). This conjugated antibody react with the human protein of CD44 and is validated by the manufacturer for flow cytometry. https://www.bdbiosciences.com/br/applications/research/t-cell-immunology/t-follicular-helper-tfh-cells/surface-markers/human/pe-cy7-mouse-anti-human-cd44-g44-26-also-known-as-c26/p/560533 FITC- or eFluor 660- conjugated CD36 (BD Bioscience, #555454 and 50-0369-42, Thermofisher). This conjugated antibody recognizes the human CD36 protein and is validated by the manufacturer for flow cytometry application. https://www.bdbiosciences.com/eu/applications/research/stem-cell-research/hematopoietic-stem-cell-markers/human/negative-markers/fitc-mouse-anti-human-cd36-cb38-also-known-as-nl07/p/555454. https://www.thermofisher.com/antibody/product/CD36-Antibody-clone-eBioNL07-NL07-Monoclonal/50-0369-42.

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)	SCC25, FaDu, and LentiX 293T were obtained from ATCC (https://www.lgcstandards-atcc.org). NHEK cell lines were obtained from Promocell (https://promocell.com/product/normal-human-epidermal-keratinocytes-nhek/). Biological samples to generate the patient-derived lines VDH01 and VDH15 were obtained from patients from the Hospital Vall d'Hebron (Barcelona, Spain) under informed consent and approval of the Bank of Tumour Committees of the hospital according to Spanish ethical regulations. The study followed the guidelines of the Declaration of Helsinki, and patient identity and pathological specimens remained anonymous in the context of the study.
Authentication	No other independent authentification was performed
Mycoplasma contamination	All cell lines were test for mycoplasma contamination on a regular basis. Cells used in this study were tested negative.
Commonly misidentified lines (See ICLAC register)	No commonly misidentified lines were used in this study

Animals and other organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research

Laboratory animals	6-8 weeks old female NSG (NOD.Cg-PrkdcSCIDIl2rgtm1Wjl/SzJ) from Jackson Laboratory (Cat#0055579) were used in this study. All

Laboratory animals	animals are kept in compliance with Annex III of EU Directive 2021/63 EU from birth to death. All mice are group housed under specific pathogen-free conditions in individually ventilated cages adapted to their body weight and behavioural biology. Food, water and nesting material are provided ad libitum. Cages are placed under a 12 hours light-dark cycle and room temperature, humidity, air change/hour, and noise is maintained and monitored. None of the mice were involved in any previous procedures before the study. Daily health checks by visual inspection were performed by the animal keepers. In case of abnormalities, individual animals are labelled as "daily observation" in the database and an automated e-mail is sent to the experimenters and the respective veterinarian/ person on site to ensure that no animal is kept beyond the humane endpoints defined in the respective animal licence.
Wild animals	No wild animals were used.
Field-collected samples	No field collected samples were used.
Ethics oversight	All mice were housed in the DKFZ Central Animal Laboratory. All mouse husbandry and experiments were carried out according to the local ethics committee (DKFZ and the regional council of Baden-Württemberg state, Regierungsprasidium Karlsruhe) under the terms and condition of the animal license G-351/19.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Human research participants

Policy information about studies involving human research participants

Population characteristics	For this study tumour sections of 78 patients of the HIPO-HNC cohort were used for IHC staining. The median age at the time of diagnose was 61.1 years (range 39.7-82.5 years), most patients were males (n=62, 79.5%) and had a history of smoking (n=57, 73.1%). A history of alcohol abuse was recorded for 41 patients (52.6%), and HPV16-related tumors (n=24, 30.8%) were almost exclusively found in the subgroup of oropharyngeal squamous cell carcinoma (OPSCC). All patients were treated by surgery and most had adjuvant radiotherapy with (n=32, 41.0%) or without platinum-based chemotherapy (n=26, 33.3%) The median follow-up period was 39.93 months (range 1.73-78.27 months).
Recruitment	Patients of the Heidelberg Center for Personalized Oncology-Head and Neck Cancer (HIPO-HNC) cohort were treated between 2012 and 2016 at the University Hospital Heidelberg, Germany, and the cohort consists primarily of advanced HNSCC without evidence of distant metastasis at the time of diagnosis.
Ethics oversight	Patient samples were obtained and analyzed under protocols S-206/2011 and S-220/2016, approved by the Ethics Committee of Heidelberg University, with written informed consent from all participants. This study was conducted in accordance with the Declaration of Helsinki.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Flow Cytometry

Plots

Confirm that:

☐ The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).

☒ The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).

☐ All plots are contour plots with outliers or pseudocolor plots.

☒ A numerical value for number of cells or percentage (with statistics) is provided.

Methodology

Sample preparation	Cells were trypsinized, washed 2 times with PBS, and fixed for 10 minutes with paraformaldehyde 1% in PBS. Following 2 additional PBS washes, cells were incubated with combinations of antibodies for 30minutes. For cell sorting, cells trypsinized, washed 2 times with PBS and incubated with combinations of antibodies for 30minutes. After incubation, cells were washed twice in PBS and analyzed or flow sorted.
Instrument	BD LSRFortessaTM analyzer or a Cell sorter (BD Biosciences).
Software	FlowJo V10 software, BD FACSDiva 8.0.
Cell population abundance	The abundance of the relevant cell populations within post-sort fractions was 80-100% in experiments.
Gating strategy	Cells were first gated regarding FSC/SCC to eliminate debris, followed by FSC-H/FSC-A gating to exclude the doublet. The markers of interest were used to identify the subpopulation to analyse. For FACS sorting, cells were selected on the basis of their forward and side scatter excluding cellular debris. Doublets and dead cells were eliminated by DAPI or propidium iodide.

☒ Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.