

# Palaeo–Eskimo genetic ancestry and the peopling of Chukotka and North America

Pavel Flegontov<sup>1,2,3\*</sup>, N. Ezgi Altınışık<sup>1,27</sup>, Piya Changmai<sup>1,27</sup>, Nadin Rohland<sup>4</sup>, Swapan Mallick<sup>4,5,6</sup>, Nicole Adamski<sup>4,5</sup>, Deborah A. Bolnick<sup>7,8</sup>, Nasreen Broomandkhoshbacht<sup>4,5</sup>, Francesca Candilio<sup>9,10</sup>, Brendan J. Culleton<sup>11</sup>, Olga Flegontova<sup>1,2</sup>, T. Max Friesen<sup>12</sup>, Choongwon Jeong<sup>13</sup>, Thomas K. Harper<sup>14</sup>, Denise Keating<sup>9</sup>, Douglas J. Kennett<sup>11,14,26</sup>, Alexander M. Kim<sup>4,15</sup>, Thiseas C. Lamnidis<sup>13</sup>, Ann Marie Lawson<sup>4,5</sup>, Iñigo Olalde<sup>4</sup>, Jonas Oppenheimer<sup>4,5</sup>, Ben A. Potter<sup>16</sup>, Jennifer Raff<sup>17</sup>, Robert A. Sattler<sup>18</sup>, Pontus Skoglund<sup>4,19</sup>, Kristin Stewardson<sup>4,5</sup>, Edward J. Vajda<sup>20</sup>, Sergey Vasilyev<sup>21</sup>, Elizaveta Veselovskaya<sup>21</sup>, M. Geoffrey Hayes<sup>22,23,24</sup>, Dennis H. O’Rourke<sup>17</sup>, Johannes Krause<sup>13</sup>, Ron Pinhasi<sup>25</sup>, David Reich<sup>4,5,6\*</sup> & Stephan Schiffels<sup>13\*</sup>

**Much of the American Arctic was first settled 5,000 years ago, by groups of people known as Palaeo-Eskimos. They were subsequently joined and largely displaced around 1,000 years ago by ancestors of the present-day Inuit and Yup’ik<sup>1–3</sup>. The genetic relationship between Palaeo-Eskimos and Native American, Inuit, Yup’ik and Aleut populations remains uncertain<sup>4–6</sup>. Here we present genomic data for 48 ancient individuals from Chukotka, East Siberia, the Aleutian Islands, Alaska, and the Canadian Arctic. We co-analyse these data with data from present-day Alaskan Iñupiat and West Siberian populations and published genomes. Using methods based on rare-allele and haplotype sharing, as well as established techniques<sup>4,7–9</sup>, we show that Palaeo-Eskimo-related ancestry is ubiquitous among people who speak Na-Dene and Eskimo–Aleut languages. We develop a comprehensive model for the Holocene peopling events of Chukotka and North America, and show that Na-Dene-speaking peoples, people of the Aleutian Islands, and Yup’ik and Inuit across the Arctic region all share ancestry from a single Palaeo-Eskimo-related Siberian source.**

Present-day Native Americans descend from at least four distinct streams of ancient migration from Asia<sup>4,5,10–12</sup>. First, people related to present-day East Asians moved into North and South America by approximately 14,500 years ago<sup>5,13,14</sup>. Here we refer to these groups as ‘First Peoples’. Second, people with a higher degree of genetic relatedness to Australasians, termed ‘Population Y’, contributed distinct ancestry to Indigenous groups from Amazonia<sup>5,10–12</sup>. Third, a stream of ancestry that relates to Palaeo-Eskimos spread throughout the American Arctic after about 5,000 years ago<sup>1–3</sup>. Fourth, a lineage that we here call ‘Neo-Eskimo’ spread with the Thule and related archaeological cultures throughout the Arctic region around 800 years ago<sup>2,3</sup>, and is today present in Yup’ik and Inuit groups. We use the terms Palaeo-Eskimo and Neo-Eskimo<sup>2,15</sup> here, but recognize that this terminology is not universally accepted by all scholars and Indigenous groups in Canada and the USA<sup>16</sup>. For naming the Arctic meta-populations, we use names of recognized language families—Na-Dene, Eskimo–Aleut, and Chukotko-Kamchatkan. We chose these terms because genetic and linguistic relationship patterns are highly congruent in this region.

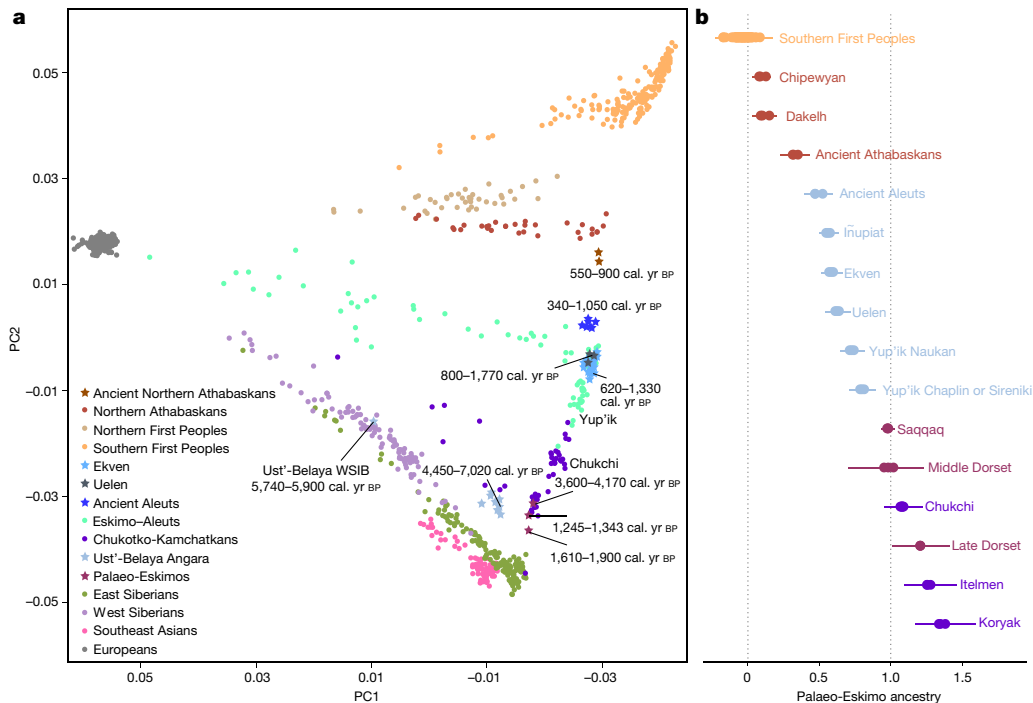
Of the four ancient sources for present-day Native Americans, the extent of Palaeo-Eskimo ancestry in living and ancient people is arguably the least understood. Although the archaeological record in the Arctic provides clear evidence for Palaeo-Eskimo cultures from about 5,000–700 years ago<sup>3,17–19</sup>, whether or not they contributed genetically to other Arctic groups is unclear. It has been argued<sup>4</sup> that Indigenous groups (such as Tlingit and Athabaskans) who speak languages of the Na-Dene family derive part of their ancestry from Palaeo-Eskimos, but other studies have challenged this finding<sup>5,6</sup>. Whether or not there was admixture between Palaeo- and Neo-Eskimos is another unresolved issue<sup>2,15,20</sup>.

We generated genome-wide data from 48 ancient individuals from the American Arctic and Siberia: 11 ancient Aleutian Islanders (dated to 2,050–280 calibrated years (cal. yr) before present (BP; taken to be AD 1950)), three ancient Northern Athabaskans (900–550 cal. yr BP), 21 individuals from the Ekven and Uelen burial grounds associated with the Chukotkan Old Bering Sea culture (1,770–620 cal. yr BP), one Palaeo-Eskimo of the Middle Dorset culture (1,900–1,610 cal. yr BP), and 12 individuals from the Ust’-Belaya burial ground near Lake Baikal (7,020–610 cal. yr BP) (Supplementary Tables 1 and 2, Supplementary Information sections 1 and 2). For each of these 48 individuals, we prepared powder from skeletal remains in a clean room, extracted DNA<sup>21</sup>, and prepared sequencing libraries, which we treated with enzymes to reduce the rate of characteristic ancient-DNA damage<sup>22</sup>. We enriched the libraries for a targeted set of approximately 1.24 million single-nucleotide polymorphisms (SNPs)<sup>23</sup>, and selected one ancient Athabaskan and one ancient Aleutian Islander for deeper shotgun sequencing (Supplementary Information section 3). In addition to these ancient data, we report SNP genotyping data for five present-day populations from Alaska and Siberia (Supplementary Table 3).

Because this study analyses DNA to understand how ancient peoples are related to present-day Indigenous peoples, we consulted with Indigenous communities in the United States and Canada regarding the study of all ancient individuals. In accordance with published guidelines for ethical genomic research with Indigenous peoples and their ancestors in the Americas<sup>24</sup>, we obtained permissions for destructive sampling of the ancient Aleuts, ancient Athabaskans, and the ancient

<sup>1</sup>Department of Biology and Ecology, Faculty of Science, University of Ostrava, Ostrava, Czech Republic. <sup>2</sup>Institute of Parasitology, Biology Centre, Czech Academy of Sciences, České Budějovice, Czech Republic. <sup>3</sup>A. A. Kharkevich Institute for Information Transmission Problems, Russian Academy of Sciences, Moscow, Russia. <sup>4</sup>Department of Genetics, Harvard Medical School, Boston, MA, USA. <sup>5</sup>Howard Hughes Medical Institute, Harvard Medical School, Boston, MA, USA. <sup>6</sup>Broad Institute of MIT and Harvard, Cambridge, MA, USA. <sup>7</sup>Department of Anthropology, University of Connecticut, Storrs, CT, USA. <sup>8</sup>Institute for Systems Genomics, University of Connecticut, Storrs, CT, USA. <sup>9</sup>School of Archaeology, University College Dublin, Dublin, Ireland. <sup>10</sup>Soprintendenza Archeologia, Belle Arti e Paesaggio per la città metropolitana di Cagliari e le province di Oristano e Sud Sardegna, Cagliari, Italy. <sup>11</sup>Institutes of Energy and the Environment, Pennsylvania State University, University Park, PA, USA. <sup>12</sup>Department of Anthropology, University of Toronto, Toronto, ON, Canada. <sup>13</sup>Department of Archaeogenetics, Max Planck Institute for the Science of Human History, Jena, Germany. <sup>14</sup>Department of Anthropology, Pennsylvania State University, University Park, PA, USA. <sup>15</sup>Department of Anthropology, Harvard University, Cambridge, MA, USA. <sup>16</sup>Department of Anthropology, University of Alaska Fairbanks, Fairbanks, AK, USA. <sup>17</sup>Department of Anthropology, University of Kansas, Lawrence, KS, USA. <sup>18</sup>Tanana Chiefs Conference, Fairbanks, AK, USA. <sup>19</sup>Francis Crick Institute, London, UK. <sup>20</sup>Department of Modern and Classical Languages, Western Washington University, Bellingham, WA, USA. <sup>21</sup>Institute of Ethnology and Anthropology, Russian Academy of Sciences, Moscow, Russia. <sup>22</sup>Department of Medicine, Northwestern University Feinberg School of Medicine, Chicago, IL, USA. <sup>23</sup>Department of Anthropology, Northwestern University, Evanston, IL, USA. <sup>24</sup>Center for Genetic Medicine, Northwestern University Feinberg School of Medicine, Chicago, IL, USA. <sup>25</sup>Department of Anthropology, University of Vienna, Vienna, Austria. <sup>26</sup>Present address: Department of Anthropology, University of California, Santa Barbara, CA, USA. <sup>27</sup>These authors contributed equally: N. Ezgi Altınışık, Piya Changmai.

\*e-mail: pavel.flegontov@osu.cz; reich@genetics.med.harvard.edu; schiffels@shh.mpg.de



**Fig. 1 | Principal component analysis and qpAdm modelling.** **a**, The first two principal components (PCs) for 940 individuals from the HumanOrigins dataset are plotted. No outliers were excluded for this analysis, which was based on 586,487 loci. Calibrated radiocarbon dates (in cal. yr BP) are shown for ancient samples (95% confidence intervals for individuals, minimal and maximal average dates for groups). See Extended Data Fig. 2 and Supplementary Information section 4 for PCA plots of additional datasets. Ust’-Belaya WSIB refers to individual I7760 from the Ust’-Belaya site, who has a West Siberian genetic profile

Middle Dorset individual, as detailed in Supplementary Information section 1. Approval was also granted for the inclusion of present-day Inupiat samples as described.

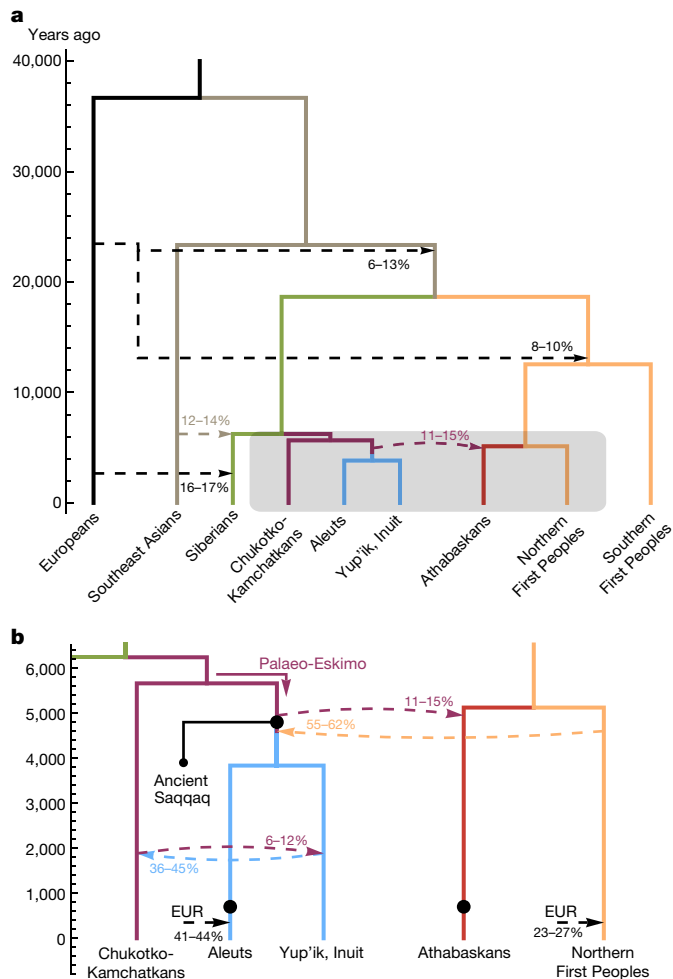
Principal component analysis (PCA) (Fig. 1a) of these data together with present-day reference data (Extended Data Fig. 1) revealed a linear cline in which Palaeo-Eskimos and some Koryaks and Itelmens (Chukotko–Kamchatkan speakers) are positioned at one extreme; then—in order—Chukchi, Yup’ik, ancient people associated with the Old Bering Sea culture and present-day Inuit, present-day and ancient Aleuts (Eskimo–Aleut speakers), ancient Athabaskans, present-day Na-Dene speakers, Northern First Peoples, and, finally, Southern First Peoples at the other extreme (Extended Data Fig. 2, Supplementary Information section 4). We used qpWave<sup>4</sup> (see Methods) to verify that this qualitative pattern in the PCA is consistent with being derived through mixture of just two sources of ancestry. When we included Chukotko–Kamchatkan speakers as target groups instead of outgroups, all of the populations on the PCA cline could be modelled as descending from two streams of ancestry (Supplementary Information section 5). We here term these two ancestry components ‘First Peoples’ and ‘proto-Palaeo-Eskimos’ (PPE). Using the qpWave extension qpAdm<sup>7</sup>, we then estimated the ancestry proportions for groupings of people along the PCA cline. Consistent with the positions on the cline, our estimates for PPE ancestry were 0% (Southern First Peoples), 0–18% (Northern First Peoples), 5–23% (present-day Na-Dene speakers), 32–43% (ancient Northern Athabaskans), 43–64% (ancient Aleuts, ancient Old Bering Sea people, and present-day Inuit), 72–82% (Yup’ik), and up to 100% in Chukotko–Kamchatkan speakers and Palaeo-Eskimos (Fig. 1b, Extended Data Figs. 3, 4). A previous analysis that used a similar setup but included Koryak in the outgroups revealed three, rather than two, lines of ancestry in Northern American populations<sup>4</sup>. We were able to reproduce this finding (Supplementary Information section 5), but as we show below,

according to PCA and ADMIXTURE analyses—different from the remaining eight individuals from the same site (labelled Ust’-Belaya Angara). **b**, Proportions of Palaeo-Eskimo ancestry inferred by qpAdm, using the same dataset as **a** but without transition polymorphisms. To visualize both systematic and statistical errors, for each target group the ancestry proportions and their single standard error intervals are shown for population triplets including different First Peoples ancestry sources, or for many alternative target groups in the case of Southern First Peoples. Target population sizes ranged from 1 to 23 individuals (5.6 on average).

the most parsimonious model for the genetic history of Chukotko–Kamchatkan speakers involves gene backflow; that is, gene flow back into Asia from Neo-Eskimos who carry both Palaeo-Eskimo and First Peoples ancestry. This backflow causes qpWave to report a separate ancestral lineage in Eskimo–Aleut speakers.

To further investigate whether the PPE source that contributes to Na-Dene-speaking people is directly related to Palaeo-Eskimos, we used ChromoPainter<sup>25</sup> to compute the cumulative length of haplotypes that are shared with the ancient Saqqaq genome<sup>1</sup>. We found that most Native American individuals with the highest relative Saqqaq-haplotype sharing belong to Na-Dene-speaking groups. This enrichment could not be explained by either Neo-Eskimo or European ancestry in these individuals (Extended Data Fig. 5, Supplementary Information section 6). Furthermore, using GLOBETROTTER<sup>26</sup>, a tool based on haplotype sharing, we identified Palaeo-Eskimos (represented by the Saqqaq individual) and First Peoples as the most probable sources of ancestry for Na-Dene speakers. Using this method, the Palaeo-Eskimo contribution ranged from 7 to 51%, and gene flow was estimated to have occurred between 2,202 and 479 years ago (Supplementary Information section 7).

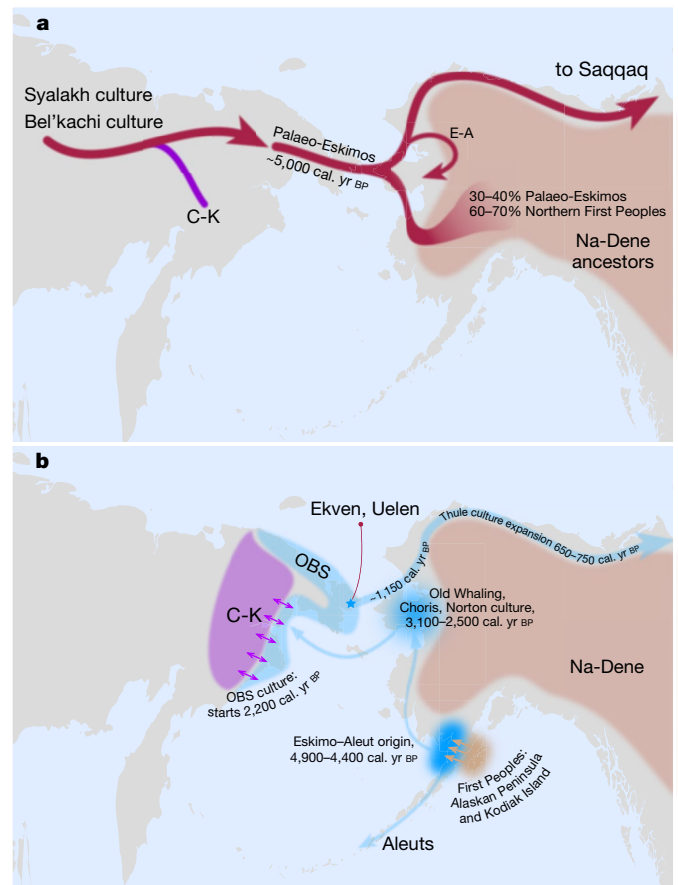
As an independent assessment of the PPE admixture cline model, we identified rare genetic variants in a large dataset of present-day full genomes outside of America and counted how often a given American genome shared those alleles. This approach allowed us to detect subtle differences in the ancestry of Indigenous populations in the Americas (Supplementary Information section 8). We found that the ancestry of present-day Athabaskans and the ancient Athabaskan and Aleut individuals with shotgun-sequenced genomes is consistent with the two-way admixture model between Palaeo-Eskimos and First Peoples—Saqqaq-related ancestry was 29–38% for present-day Athabaskans, approximately 42% for the ancient Athabaskan, and approximately 65% for the ancient Aleut (Extended Data Fig. 6).



**Fig. 2 | A demographic model based on 114 individuals from 9 meta-populations. a**, We used Rarecoal and qpGraph to test topologies and estimate split times and admixture edges (dashed arrows). For a complete list of parameter estimates, including confidence intervals, see Supplementary Information section 9. **b**, A zoomed-in model for the past 6,000 years and for 5 populations, highlighting the Holocene migrations and gene-flow events between Asia and America. Maximum-likelihood branching points of the ancient Saqqaq, Aleut and Athabaskan genomes are indicated as solid dots on internal branches. The 11–15% Palaeo-Eskimo admixture proportion into the ancestors of present-day Athabaskans is less than the approximately 32–43% estimated from Ancient Athabaskans, reflecting reduction of Palaeo-Eskimo ancestry through later admixtures with northern First Peoples. Times are scaled using a per-generation mutation rate<sup>28</sup> of  $1.25 \times 10^{-8}$  and a generation time of 29 years<sup>29</sup> (see Supplementary Information section 9). EUR, Europeans.

In this analysis, the qpAdm analysis, and further analyses below we obtained a proportion of PPE ancestry that was consistently higher in ancient compared with present-day Athabaskans. This suggests that ongoing bidirectional genetic exchange with neighbouring Northern First Peoples has been reducing PPE ancestry in Na-Dene-speaking people. Rare-allele sharing also shows that present-day Yup'ik and Inuit genomes are inconsistent with this two-way admixture model, and instead exhibit higher allele sharing with the genomes of Chukotko-Kamchatkan speakers. This is consistent with the qpWave and qpAdm analysis above, and with our explicit demographic model below.

We used qpGraph to build a demographic model for the populations analysed here (Supplementary Information section 10). To explore the model space as far as possible, at each stage of its development we kept all fitting models that connected a given set of populations. We explicitly tested all possible topologies within the PPE clade, which



**Fig. 3 | Archaeological and geographical interpretation of our model. a**, The topology drawn here reflects our best-fitting model of the PPE clade. We provisionally mapped the gene flow from Palaeo-Eskimos to Na-Dene speakers across the boundary that separates the Arctic Small Tool and Northern Archaic traditions in Alaska. This is where the highest diversity of Na-Dene languages is found today (for that reason, Alaska was proposed as a Na-Dene homeland<sup>30</sup>). C-K, Chukotko-Kamchatkan speakers; E-A, Eskimo-Aleut speakers. **b**, A model of population history for Eskimo-Aleut speakers, combining genetic and archaeological evidence. Their back-and-forth movement across the Bering Strait is illustrated, as well as the bidirectional gene flow between Yup'ik and Inuit ancestors (the Old Bering Sea culture, OBS) and Chukotko-Kamchatkan speakers in Chukotka. In both panels, the earliest dates in cal. yr BP are indicated for archaeological areas and migrations. Some migration paths are drawn to indicate the general directions, but not the actual routes, of population spread.

consists of Chukotko-Kamchatkan speakers, Eskimo-Aleut speakers, Athabaskans, and the ancient Saqqaq individual. After testing 224 models, we found that the best-fitting topology of this clade had a grouping (C-K, (ATH<sub>PPE</sub>, (SAQ, E-A<sub>PPE</sub>))) (Extended Data Fig. 7), in which C-K represents Chukotko-Kamchatkan speakers, ATH<sub>PPE</sub> the PPE source in Athabaskans, SAQ the Saqqaq genome, and E-A<sub>PPE</sub> the PPE source in Eskimo-Aleut speakers. The C-K group splits off before the PPE source in Athabaskans. A key feature in our best-fitting model is bidirectional gene flow, which occurred between Chukotko-Kamchatkan-speaking and Neo-Eskimo populations but did not affect Aleuts—consistent with the qpWave and rare-allele-sharing analyses. We further investigated the population history of the Aleuts—who can be grouped by burial tradition into Neo- and Palaeo-Aleuts—by analysing our ancient Aleut genomes, and found that according to PCA (Fig. 1a), ADMIXTURE (Extended Data Fig. 8), and allele-sharing analyses (Supplementary Information section 11), these two groups are consistent with one genetically homogenous population, contradicting previous suggestions for movements of new people into the Aleutian islands around 1000 cal. yr BP<sup>27</sup>.



We then used Rarecoal to test the final graph topology that we obtained using qpGraph, and to infer split times (Supplementary Information section 9). Our final model (Fig. 2) suggests that the Chukotko-Kamchatkan and Eskimo–Aleut lineages diverged 4,900–6,200 years ago; that the time of the PPE gene flow into the ancestors of Athabaskans was 4,400–5,000 years ago; and that the branch position of the Saqqaq individual is immediately after the latter gene flow. We also find that interactions with Northern First Peoples around 4,400–4,900 years ago (consistent with estimates from the ALDER method; Supplementary Information section 12) resulted in this group contributing 55–62% genetic ancestry to ancestors of Eskimo–Aleut-speaking populations. Finally, we estimate that the time of bidirectional gene flow between the Chukotko-Kamchatkan and Eskimo–Aleut lineages was 1,700–2,300 years ago (6–15% Chukotko-Kamchatkan contribution to Eskimo–Aleut, and 36–45% Eskimo–Aleut to Chukotko-Kamchatkan; but see lower estimates in the qpGraph model, Extended Data Fig. 7). Our final model also contains substantial European gene flow (presumably during the colonial period) into present-day Aleuts (approximately 41–44%) and Northern First Peoples (approximately 23–27%). We note that our best-fitting topology differs from a previously published model with a PPE grouping of the form ((C-K, ATH<sub>PPE</sub>), (SAQ, E-A<sub>PPE</sub>)), in which Chukotko-Kamchatkans and the PPE source in Athabaskans are sister clades<sup>6</sup>. We compared this model and other topologies with ours, and found that our proposed topology was a significantly better fit, according to various qpGraph metrics and the substantial likelihood differences that were reported by Rarecoal (see Supplementary Information sections 9 and 10 for a description of statistical tests). A previous model<sup>6</sup> found no evidence for ancestry related to an 11,500-year-old individual from Alaska in Athabaskans (figure 3 of that study, although see a contradicting model in Supplementary Section 18 of the same study<sup>6</sup>), and when we explicitly tested this using qpGraph, we found that our model supported this conclusion.

Genetic data can document the existence and timing of interactions such as the ones that gave rise to the ancestors of Eskimo–Aleut and Na-Dene speakers, but without ancient DNA that originates from the specific times and places that these interactions occurred, it is impossible to pinpoint their geographic location. On the basis of archaeological evidence and parsimony, however, the most plausible scenario is that the gene-flow events giving rise to Eskimo–Aleut and to Na-Dene speakers occurred in Alaska (Fig. 3)—we discuss the archaeological and linguistic implications of this model in the Supplementary Discussion and in Supplementary Information section 13. A priority for future work should be to analyse samples from Alaska that date to our proposed time windows of admixture in the third millennium BCE.

**Note added in proof:** After this manuscript was accepted, another study was published<sup>31</sup>, which also analyses the newly reported genotyping data from the 35 Inupiat individuals. Our manuscript and ref.<sup>31</sup> both provide details of the data generation and data access, and either can be cited for the publication of these data.

## Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at <https://doi.org/10.1038/s41586-019-1251-y>.

Received: 9 October 2017; Accepted: 25 February 2019;  
Published online: 05 June 2019

- Rasmussen, M. et al. Ancient human genome sequence of an extinct Palaeo-Eskimo. *Nature* **463**, 757–762 (2010).
- Raghavan, M. et al. The genetic prehistory of the New World Arctic. *Science* **345**, 1255832 (2014).
- Friesen, T. M. in *The Oxford Handbook of the Prehistoric Arctic* (eds Friesen, T. M. & Mason, O. K.) 673–692 (Oxford Univ. Press, New York, 2016).
- Reich, D. et al. Reconstructing Native American population history. *Nature* **488**, 370–374 (2012).
- Raghavan, M. et al. Genomic evidence for the Pleistocene and recent population history of Native Americans. *Science* **349**, aab3884 (2015).

- Moreno-Mayar, J. V. et al. Terminal Pleistocene Alaskan genome reveals first founding population of Native Americans. *Nature* **553**, 203–207 (2018).
- Haak, W. et al. Massive migration from the steppe was a source for Indo-European languages in Europe. *Nature* **522**, 207–211 (2015).
- Patterson, N. et al. Ancient admixture in human history. *Genetics* **192**, 1065–1093 (2012).
- Schiffels, S. et al. Iron Age and Anglo-Saxon genomes from East England reveal British migration history. *Nat. Commun.* **7**, 10408 (2016).
- Skoglund, P. et al. Genetic evidence for two founding populations of the Americas. *Nature* **525**, 104–108 (2015).
- Moreno-Mayar, J. V. et al. Early human dispersals within the Americas. *Science* **362**, eaav2621 (2018).
- Posth, C. et al. Reconstructing the deep population history of Central and South America. *Cell* **175**, 1185–1197.e22 (2018).
- Potter, B. A. et al. Early colonization of Beringia and Northern North America: chronology, routes, and adaptive strategies. *Quat. Int.* **444**, 36–55 (2017).
- Llomas, B. et al. Ancient mitochondrial DNA provides high-resolution time scale of the peopling of the Americas. *Sci. Adv.* **2**, e1501385 (2016).
- Raff, J. A., Rzhetskaya, M., Tackney, J. & Hayes, M. G. Mitochondrial diversity of Inupiat people from the Alaskan North Slope provides evidence for the origins of the Paleo- and Neo-Eskimo peoples. *Am. J. Phys. Anthropol.* **157**, 603–614 (2015).
- Friesen, T. M. On the naming of Arctic archaeological traditions: the case for Paleo-Inuit. *Arctic* **68**, iii–iv (2015).
- Park, R. W. in *The Oxford Handbook of the Prehistoric Arctic* (eds Friesen, T. M. & Mason, O. K.) 417–442 (Oxford Univ. Press, New York, 2016).
- Prentiss, A. M., Walsh, M. J., Foor, T. A. & Barnett, K. D. Cultural macroevolution among high latitude hunter-gatherers: a phylogenetic study of the Arctic Small Tool tradition. *J. Archaeol. Sci.* **59**, 64–79 (2015).
- Tremayne, A. H. & Rasic, J. T. in *The Oxford Handbook of the Prehistoric Arctic* (eds Friesen, T. M. & Mason, O. K.) 303–322 (Oxford Univ. Press, New York, 2016).
- Friesen, T. M. Contemporaneity of Dorset and Thule cultures in the North American Arctic: new radiocarbon dates from Victoria Island, Nunavut. *Curr. Anthropol.* **45**, 685–691 (2004).
- Dabney, J. et al. Complete mitochondrial genome sequence of a Middle Pleistocene cave bear reconstructed from ultrashort DNA fragments. *Proc. Natl Acad. Sci. USA* **110**, 15758–15763 (2013).
- Rohland, N., Harney, E., Mallick, S., Nordenfelt, S. & Reich, D. Partial uracil-DNA-glycosylase treatment for screening of ancient DNA. *Phil. Trans. R. Soc. Lond. B* **370**, 20130624 (2015).
- Fu, Q. et al. An early modern human from Romania with a recent Neanderthal ancestor. *Nature* **524**, 216–219 (2015).
- Bardill, J. et al. Advancing the ethics of paleogenomics. *Science* **360**, 384–385 (2018).
- Lawson, D. J., Hellenthal, G., Myers, S. & Falush, D. Inference of population structure using dense haplotype data. *PLoS Genet.* **8**, e1002453 (2012).
- Hellenthal, G. et al. A genetic atlas of human admixture history. *Science* **343**, 747–751 (2014).
- Smith, S. E. et al. Inferring population continuity versus replacement with aDNA: a cautionary tale from the Aleutian Islands. *Human Biol.* **81**, 407–426 (2009).
- Scally, A. & Durbin, R. Revising the human mutation rate: implications for understanding human evolution. *Nat. Rev. Genet.* **13**, 745–753 (2012).
- Fenner, J. N. Cross-cultural estimation of the human generation interval for use in genetics-based population divergence studies. *Am. J. Phys. Anthropol.* **128**, 415–423 (2005).
- Kari, J. in *The Dene-Yeniseian Connection, Anthropological Papers of the University of Alaska: New Series*, vol. 5 (eds Kari, J. & Potter, B. A.) 194–222 (Univ. of Alaska and Alaska Native Language Centre, Fairbanks, Alaska, 2010).
- Reynolds, A. W. et al. Comparing signals of natural selection between three Indigenous North American populations. *Proc. Natl Acad. Sci. USA* **116**, 9312–9317.

**Acknowledgements** We acknowledge the ancient people whose skeletal samples were studied, the Aleut Corporation, the Aleutians Pribilof Islands Association, and the Chaluka Corporation for granting permissions to conduct genetic analyses on the eastern Aleutians. We thank the staff at the Smithsonian Institution's National Museum of Natural History for facilitating the sample collection; the McGrath Native Village Council and MTNT Ltd for granting permissions to conduct genetic analyses on the Tochak McGrath remains; J. Clark, who performed biological age estimates on these remains; the research participants in Alaska (Genetics of Alaskan North Slope (GeANS) project funded by NSF OPP-0732857) and West Siberia who donated samples for genome-wide analysis; J. B. Coltrain for sharing data on stable isotopes; and J. W. Ives, J. Tackney, L. Norman, and K. TallBear for comments on earlier drafts of this paper. Sample collection and the initial molecular, isotopic, and accelerator mass spectrometry (AMS) <sup>14</sup>C dating of the samples described here were funded by National Science Foundation Office of Polar Program grants OPP-9726126, OPP-9974623, and OPP-0327641; the Natural Sciences and Engineering Research Council of Canada; and the Wenner-Gren Foundation for Anthropological Research (6364). This work was supported by the Czech Ministry of Education, Youth and Sports from the project 'IT4Innovations National Supercomputing Center – LM2015070'. P.F., P.C., O.F., and N.E.A. were supported by the Institutional Development Program of the University of Ostrava; P.F. and P.C. were supported by the EU Operational Programme 'Research and Development for Innovations' (CZ.1.05/2.1.00/19.0388) and P.C. was also supported by the Statutory City of Ostrava (0924/2016/ŠaS)



and the Moravian-Silesian Region (O1211/2016/RRC); P.S. was funded by the Francis Crick Institute, which receives its core funding from Cancer Research UK (FC001595), the UK Medical Research Council (FC001595), and the Wellcome Trust (FC001595); D.R. was funded by NSF HOMINID (grant BCS-1032255), NIH (NIGMS), the Allen Discovery Center of the Paul Allen Foundation (grant GM100233), and is an Investigator of the Howard Hughes Medical Institute; D.A.B. was supported by a Norman Hackerman Advanced Research Program grant from the Texas Higher Education Coordinating Board; AMS  $^{14}\text{C}$  work at Pennsylvania State University by D.J.K. and B.J.C. was funded by the NSF Archaeometry programme (BCS-1460369); and C.J., T.C.L., J.K., and S.S. were supported by the Max Planck Society.

**Reviewer information** *Nature* thanks Carles Lalueza-Fox, John Lindo and the other anonymous reviewer(s) for their contribution to the peer review of this work.

**Author contributions** S.S., P.F., and D.R. supervised the study. B.A.P., T.M.F., A.M.K., R.A.S., S.V., E.V., D.H.O'R., R.P., and D.R. assembled the collection of archaeological samples. D.A.B., O.F., J.R., M.G.H., and J.K. assembled the sample collection from present-day populations. T.K.H., D.J.K., B.J.C., and T.M.F. were responsible for radiocarbon dating and calibration. N.R., N.A., N.B.,

F.C., D.K., A.M.L., J.O., and K.S. performed laboratory work and supervised ancient DNA sequencing. P.F., N.E.A., P.C., S.M., C.J., T.C.L., I.O., P.S., and S.S., analysed genetic data. E.J.V. wrote the supplemental section on linguistics. P.F., D.R., and S.S. wrote the manuscript with additional input from all other co-authors.

**Competing interests** The authors declare no competing interests.

**Additional information**

**Extended data** is available for this paper at <https://doi.org/10.1038/s41586-019-1251-y>.

**Supplementary information** is available for this paper at <https://doi.org/10.1038/s41586-019-1251-y>.

**Reprints and permissions information** is available at <http://www.nature.com/reprints>.

**Correspondence and requests for materials** should be addressed to P.F., D.R. or S.S.

**Publisher's note:** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2019

## METHODS

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

**Ancient DNA sampling, extraction, and sequencing.** In dedicated clean rooms at Harvard Medical School (for the 11 Aleutian Islanders, 3 Tochak McGrath samples, and 1 Middle Dorset sample), and at University College Dublin (for the 33 Chukotkan and Baikal region samples), we prepared powder from human skeletal remains, as previously described<sup>7</sup>. We extracted DNA using a previously published method<sup>21</sup>, and prepared double-stranded barcoded libraries. These were treated with uracil-DNA glycosylase to remove characteristic cytosine-to-thymine damage in ancient DNA, using a previous protocol<sup>22</sup>. We enriched the libraries for a set of approximately 1.24 million SNPs<sup>23</sup>, and sequenced on an Illumina NextSeq instrument using 75-nucleotide paired-end reads, which we merged before mapping to the human reference genome version hg19 (requiring at least 15 base pairs of overlap) (Supplementary Information section 3). We also carried out shotgun sequencing of one ancient Aleutian Islander individual and one ancient Athabaskan individual (Supplementary Table 1). The work with the ancient Native American individuals was conducted after consultation with local communities and authorities, and after formal permissions were granted. Results have been communicated in person and in writing to descendant communities.

**Sampling present-day populations.** Sampling of the Alaskan Inupiat population (35 individuals) was performed with informed consent as described<sup>15</sup> (see also Supplementary Information section 1). Saliva samples of four West Siberian ethnic groups (Enets, Kets, Nganasans, and Selkups; 58 individuals in total) were collected and DNA extractions were performed as described<sup>32</sup> (see also Supplementary Table 3). In the case of the genotyped West Siberians, the study was approved by the ethical committee of the Lomonosov Moscow State University. All volunteers signed informed consent forms. The study was also approved by local administrations of the Taymyr and Turukhansk districts and discussed with local committees of small Siberian nations for observance of their rights and traditions. In the case of the Inupiat, the study was approved by Northwestern University's Institutional Review Board, after consultation with the Ukpavik Inupiat Corporation, the Native Village of Barrow, and the Senior Advisory Council of Barrow (Elders). Study participants gave informed consent (see Supplementary Information section 1).

**Preparation of ancient genomic datasets.** We made two types of genotype calls for ancient samples. First, for merging with the 1240K SNP-capture dataset (that is, the dataset of 1.24 million SNPs) subsequently used for the qpGraph analysis, and for merging with the HumanOrigins and Illumina SNP-array datasets, we made pseudo-haploid calls using a single randomly sampled sequence at each captured position. Second, for rare-variant analysis (rare-allele-sharing statistics (RASS) and Rarecoal) we used only shotgun genomes (not exposed to SNP capture), and generated pseudo-haploid calls using the majority allele at sites that were covered by at least three sequences. This ensures that all calls are supported by at least two sequences, thus reducing the error rate. Sites covered by more than three sequences were first downsampled to three sequences, to reduce a subtle reference bias associated with the majority calling method for high coverage data. The majority call method with downsampling is implemented in the program pileupCaller, which is available at <https://www.github.com/stschiff/sequenceTools>.

**Dataset preparation for present-day genomes.** To analyse patterns of rare-allele sharing, we compiled a set of shotgun-sequencing data that covered Africa, Europe, Southeast Asia, Siberia, and the Americas. The data encompass 190 individuals from 87 populations, and include two shotgun genomes generated in this study (Supplementary Table 4). We assembled the dataset using two published sources: the Simons Genome Diversity Project<sup>33</sup> and the modern genomes published in ref.<sup>5</sup>. We used the variant calls that were generated in these publications, keeping only biallelic autosomal SNPs that were covered in at least 90% of individuals in the respective datasets. Finally, we filtered out SNPs excluded by our mappability mask, which was generated as described<sup>34</sup>, and selected populations for the rare-allele-sharing and Rarecoal analyses as described in Supplementary Information sections 8 and 9, respectively. We also compiled another dataset by overlapping this genomic dataset with the SNP-capture data at up to 1.24 million sites that we generated for ancient samples (Supplementary Table 1), and added pseudo-haploid data for the USR1<sup>6</sup>, Saqqaq<sup>1</sup>, Clovis<sup>35</sup>, MA1<sup>36</sup>, and Loschbour<sup>37</sup> ancient individuals. We then selected populations for the qpGraph analysis as described in Supplementary Information section 10. Individual, population, and site counts and filtration settings for these datasets are presented in Supplementary Table 5.

We also assembled two independent SNP-array datasets (see dataset compositions in Supplementary Table 4 and filter settings in Supplementary Table 5). First, we obtained phased autosomal genotypes for large worldwide collections of Affymetrix HumanOrigins (3,246 individuals) or Illumina (2,325 individuals) SNP-array data (Supplementary Table 5), using ShapeIt v.2.20 with default parameters and without a guidance haplotype panel<sup>38</sup>. Then we applied missing-rate

thresholds for individuals (below 50%) and SNPs (below 5%) using PLINK v.1.90b3.36<sup>39</sup>. For ADMIXTURE<sup>40</sup>, PCA, and qpWave or qpAdm<sup>4,7</sup> analyses, phasing was not performed, and more-relaxed missing-rate thresholds for ancient individuals were applied (75% or 70% depending on the dataset) (Supplementary Table 5). As a result, ancient individuals who have more than 350,000 SNP sites genotyped on the 1240K panel were selected (Supplementary Table 1). This allowed us to include relevant ancient samples that were genotyped using the targeted enrichment approach. The Middle Dorset Palaeo-Eskimo individual was included despite having a higher missing rate of 89–90% (depending on the dataset). For the ADMIXTURE analysis, unlinked SNPs were selected using linkage disequilibrium filtering with PLINK (Supplementary Table 5).

In the SNP datasets, we removed outliers manually, considering the results of an unsupervised ADMIXTURE<sup>40</sup> analysis ( $K = 14$  or  $11$  in the case of the HumanOrigins or Illumina datasets, respectively) and weighted Euclidean distances. In ADMIXTURE, we inspected individuals for non-typical ancestry components (for example, European in Native Americans). For the latter criterion, ten principal components were computed using PLINK v.1.90b3.36, and weighted Euclidean distances defined as

$$d(q, p) = \sqrt{\frac{1}{\sum_{i=1}^{10} \lambda_i} \sum_{i=1}^{10} \lambda_i (q_i - p_i)^2}$$

were calculated among individuals within populations ( $q_i$  and  $p_i$  refer to PCs from 1 to 10 in a population,  $\lambda_i$  is the corresponding eigenvalue). Individuals were identified as outliers if they had average weighted Euclidean distances from all other individuals in a population that were larger than [third quartile +  $1.5 \times$  (third quartile – first quartile)]. Manual removal of outliers based on ADMIXTURE profiles—that is, based on outstanding proportions of European and other non-typical ancestry components—was prioritized, and some individuals identified as outliers based on average weighted Euclidean distances were kept if they had a typical ADMIXTURE profile (see examples for the Ket, Nganasan, Tubalar, and Yup'ik Chaplin or Sireniki populations in the HumanOrigins dataset, Supplementary Information section 4). If most individuals in a population had European admixture, we removed only those that had the most extreme admixture proportions, to keep the final population size reasonably large (see examples for the Splatsin, Stswecem'c, Tlingit and other groups in the Illumina dataset, Supplementary Information section 4). Removal of outliers based on average weighted Euclidean distances was prioritized if all individuals had a uniform ADMIXTURE profile (see examples for the Karitiana, Mansi, Surui, Xavante, and Zapotec populations in the HumanOrigins dataset, Supplementary Information section 4). The ADMIXTURE results, Euclidean distances, PC1 versus PC2 plots, and outcomes of the outlier removal procedure for American and Siberian populations are presented in Supplementary Information section 4. We note that this outlier removal procedure preceded ChromoPainter v.1<sup>25</sup> and v.2<sup>26</sup>, fineSTRUCTURE<sup>25</sup>, haplotype-sharing statistic (HSS), and GLOBETROTTER<sup>26</sup> analyses, and the ADMIXTURE<sup>40</sup> analyses that are presented in Extended Data Fig. 8.

In the case of some analyses that relied on the Illumina SNP-array dataset (ChromoPainter v.1, HSS), Na-Dene-speaking populations were exempt from the first round of outlier removal and from removal of supposed relatives identified in ref.<sup>5</sup>. This was done to preserve maximal diversity of Na-Dene speakers and to ensure that both Dakelh individuals with sequencing data available would be included. The exemption was applied only to analyses that operate on individuals independently. Outlier removal was also not applied to the whole-genome datasets used in the RASS and Rarecoal analyses.

For the qpWave<sup>4</sup>, qpAdm<sup>7</sup>, qpGraph<sup>8</sup>, ALDER<sup>41</sup>, and  $f_4$ -statistic<sup>8</sup> analyses the first round of outlier removal was followed by a more stringent procedure. Any Native American individual with a proportion of European, African, or Southeast Asian ancestry that exceeded 1% according to ADMIXTURE (Extended Data Fig. 8) was removed, as well as Chukotkan and Kamchatkan individuals with greater than 1% European ancestry. Some additional Chipewyan and West Greenlandic Inuit individuals were removed, as European ancestry that was undetectable with ADMIXTURE was revealed in them using statistics  $D$ (Yoruba or Dai, Icelander; Chipewyan individual, Karitiana) and  $D$ (Yoruba or Dai, Slovak; West Greenlandic Inuit individual, Karitiana). Any individual for which either of the two absolute  $Z$ -scores was greater than three was removed. The outcome of the multi-step dataset pruning procedure that preceded the qpWave or qpAdm,  $f_4$ -statistic, and ALDER analyses is illustrated by the pairs of PCA plots presented in Fig. 1a, Extended Data Fig. 2 and Supplementary Information section 4.

For some analyses, we combined groups into meta-populations, as indicated in Extended Data Fig. 1 and summarized in Supplementary Table 4. The breakdown of groups into these meta-populations was guided by unsupervised clustering using ADMIXTURE (Extended Data Fig. 8), fineSTRUCTURE (Extended Data Fig. 9), PCA (Fig. 1a, Extended Data Fig. 2, Supplementary Information section 4) and by contextual information in some cases. For naming the Arctic meta-populations,

we use names of recognized language families—Na-Dene, Eskimo–Aleut, and Chukotko-Kamchatkan.

Finally, we selected relevant meta-populations and generated datasets of 489–1,184 individuals, which we analysed further with ADMIXTURE<sup>40</sup>, PCA (as implemented in PLINK v.1.90b3.36)<sup>39</sup>, qpWave or qpAdm<sup>47</sup>, ALDER<sup>41</sup>, ChromoPainter v.1 and fineSTRUCTURE<sup>25</sup>, and ChromoPainter v.2 and GLOBETROTTER<sup>26</sup> (Supplementary Tables 4, 5). Populations having, on average, more than 5% of the Siberian ancestral component according to ADMIXTURE analysis (Extended Data Fig. 8)—for example, Finns and Russians—were excluded from the European and Southeast Asian meta-populations.

To test whether the datasets used in this study would allow us to detect substructure in the First Peoples and American Arctic populations, we divided each American population that consisted of two or more individuals into two halves (equal, if possible) randomly and calculated the following  $f_4$  statistics: (American<sub>*i*</sub> half A, American<sub>*j*</sub>; American<sub>*i*</sub> half B, Dai). We show the Z scores for these statistics (Supplementary Table 6), and conclude that six dataset versions (HumanOrigins, 1240K, Illumina, with or without transition polymorphisms) have the power to distinguish American populations from each other. Population halves were matched correctly in 89–98% of cases; that is, the  $f_4$  statistics were significantly positive ( $Z > 3$ ).

**ADMIXTURE analysis.** The ADMIXTURE software<sup>40</sup> implements a model-based Bayesian approach that uses a block-relaxation algorithm to compute a matrix of ancestral-population fractions in each individual ( $Q$ ) and infer allele frequencies for each ancestral population ( $P$ ). A given dataset is usually modelled using various numbers of ancestral populations ( $K$ ). We ran ADMIXTURE v.1.23 for the HumanOrigins-based and Illumina-based datasets of unlinked SNPs (Supplementary Table 5) using  $K$  values of 10–25 and 5–20, respectively. One hundred analysis iterations were generated with different random seeds. The best run was chosen according to the highest likelihood. An optimal value of  $K$  was selected using tenfold cross-validation.

**PCA.** PCA was performed using PLINK v.1.90b3.36<sup>39</sup> with default settings. No pruning of linked SNPs was applied before this analysis (Supplementary Table 5), and almost identical results were obtained for pruned datasets.

**Admixture modelling with qpWave and qpAdm.** We used the qpWave v.310 tool (part of AdmixTools v.4.1) to infer how many streams of ancestry relate a set of test populations to a set of outgroups<sup>4</sup>. qpWave relies on a matrix of statistics,  $f_4(\text{test}_i, \text{test}_j; \text{outgroup}_1, \text{outgroup}_2)$ . Usually, a few test populations from a certain region and a diverse worldwide set of outgroups (having no recent gene flow from the test region) are co-analysed<sup>7,10,42</sup>, and a statistical test is performed to determine whether allele frequencies in the test populations can be explained by one, two, or more streams of ancestry derived from the outgroups. If a group of three populations—a triplet—is derived from two ancestry streams according to a qpWave test, and any pair of the constituent populations shows the same result, it follows that one of the populations can be modelled as having ancestry from the other two using another tool, qpAdm v.401<sup>7</sup>.

The following sets of outgroup populations were used for analyses on the HumanOrigins dataset: (1) ‘OG19’, 19 outgroups from five broad geographical regions: Mbuti, Taa, Yoruba (Africans), Nganasan, Tuvinian, Ulchi, Yakut (East Siberians), Altaian, Ket, Selkup, Tubalar (West Siberians), Czech, English, French, North Italian (Europeans), and Dai, Miao, She, Thai (Southeast Asians); (2) ‘OG19\_UB1526’, OG19 and an ancient Siberian individual I1526 (the highest-coverage individual at the Ust’-Belaya Angara site) that was distinct from the other Siberians according to our PCA analyses (Fig. 1a) and thus might increase the diversity of Siberian outgroups and the resolution of the method; (3) ‘OGA’, eight diverse Siberian populations (Nganasan, Tuvinian, Ulchi, Yakut, Even, Ket, Selkup, Tubalar) and a Southeast Asian population (Dai); (4) ‘OGA\_Koryak’, OGA and Koryak, a Chukotko-Kamchatkan-speaking group that supposedly provides higher resolution as it is closely related to the putative PPE admixture partners (Supplementary Information section 10); and (5) ‘OGA\_UB1526’, OGA and the Ust’-Belaya Angara individual I1526.

Similar sets of outgroup populations were used for analyses on the Illumina dataset: (1) ‘OG20’: Bantu (Kenya), Mandenka, Mbuti, Yoruba (Africans), Buryat, Evenk, Nganasan, Tuvinian, Yakut (East Siberians), Altaian, Khakas, Selkup (West Siberians), Basque, Sardinian, Slovak, Spanish (Europeans), and Dai, Lahu, Miao, She (Southeast Asians); (2) ‘OG20\_UB1526’, OG20 and the highest-coverage Ust’-Belaya Angara individual I1526; (3) ‘OGA’, nine Siberian populations (Buryat, Dolgan, Evenk, Nganasan, Tuvinian, Yakut, Altaian, Khakas, Selkup) and Dai; (4) ‘OGA\_Koryak’, OGA and Koryak; and (5) ‘OGA\_UB1526’, OGA and the Ust’-Belaya Angara individual I1526.

All possible triplets of the form (First Peoples or Na-Dene-speaking population; Eskimo–Aleut population; Palaeo-Eskimo or Chukotko-Kamchatkan-speaking population) and quadruplets of the form (First Peoples population; Na-Dene-speaking population; Eskimo–Aleut population; Palaeo-Eskimo or Chukotko-Kamchatkan-speaking population) were tested with qpWave for both

the HumanOrigins and Illumina SNP-array datasets, with or without transition polymorphisms, and using five alternative outgroup sets. The Koryak outgroup was not tested for population triplets or quadruplets including Chukotko-Kamchatkan speakers, as such models are expected to be non-fitting by default. For admixture inference with qpAdm, all possible triplets of the form (any American, Chukotkan or Kamchatkan population; Palaeo-Eskimo or Chukotko-Kamchatkan-speaking population; Guarani, Karitiana, or Mixe) were considered in the case of the HumanOrigins dataset, and all possible triplets of the form (any American, Chukotkan or Kamchatkan population; Palaeo-Eskimo or Chukotko-Kamchatkan-speaking population; Karitiana, Mixtec, Nisga’a, or Pima) were considered in the case of the Illumina dataset. Palaeo-Eskimos were represented by the Saqqaq (around 3,900 cal. yr BP), Middle Dorset (around 1,750 cal. yr BP), and Late Dorset individuals (around 750 cal. yr BP)—widely separated in space and time—and two types of SNP calls were tested for the Saqqaq individual: published diploid calls<sup>2</sup> with 50–58% missing rates (in various dataset versions) and pseudo-haploid calls with much lower missing rates of 4–11% (in various dataset versions) generated by us. See further details in Supplementary Information section 5.

**fineSTRUCTURE clustering.** We used fineSTRUCTURE v.2.0.7 with default parameters to analyse the output of ChromoPainter v.1<sup>25</sup>. Clustering trees of individuals were generated by fineSTRUCTURE based on counts of shared haplotypes<sup>25</sup>, and two independent iterations of the clustering algorithm were performed. The clustering trees and co-ancestry matrices were visualized using fineSTRUCTURE GUI v.0.1.0<sup>25</sup>.

**Haplotype-sharing statistics.** The haplotype-sharing statistic ( $HSS_{AB}$ ) is defined as the total genetic length of DNA (in cM) that a given individual  $A$  shares with individual  $B_j$  under the model<sup>25,26</sup>.  $HSS_{AB}$  was computed in the all-versus-all manner using ChromoPainter v.1<sup>25</sup> with default parameters. In practice, we summed up the length of DNA that individual  $A$  copied from individual  $B_j$  and the length of DNA copied in the opposite direction (from  $B_j$  to  $A$ ); that is, we disregarded the distinction between donor and recipient that was introduced by the ChromoPainter software. For each individual  $A$  (in practice, an American individual),  $HSS_{AB}$  values were averaged across all individuals of a reference population  $B$  (the Siberian or Arctic meta-population, or the Saqqaq ancient genome<sup>1</sup>), and then normalized by the haplotype-sharing statistic  $HSS_{AC}$  for the European, African, or Siberian outgroup  $C$ . The resulting statistics  $HSS_{AB}/HSS_{AC}$  are referred to as Siberian, Arctic, or Saqqaq relative haplotype sharing, and were visualized for separate individuals. Similar statistics were calculated for Siberian and Arctic individuals using the leave-one-out procedure. Relative HSSs for recently admixed populations, with ancestry from population  $A$  and population  $B$ , were calculated in the following way:  $a \times HSS_{AC}/HSS_{AD} + b \times HSS_{BC}/HSS_{BD}$ , in which  $a$  and  $b$  are admixture proportions being simulated in steps of 5%. See further details in Supplementary Information section 6.

**Dating admixture events using haplotype-sharing statistics.** We used GLOBETROTTER<sup>26</sup> (version from 27 May 2016) to infer and date up to two admixture events in the history of Na-Dene-speaking populations. To detect subtle signals of admixture between closely related source populations, we followed a previously published ‘regional’ analysis protocol<sup>26</sup>. Using ChromoPainter v.2<sup>26</sup>, chromosomes of a target Na-Dene-speaking population were ‘painted’ as a mosaic of haplotypes derived from donor populations or meta-populations: the Saqqaq ancient genome, Chukotko-Kamchatkan- and Eskimo–Aleut-speaking groups, Northern First Peoples, Southern First Peoples, West Siberians, East Siberians, Southeast Asians, and Europeans. Target individuals were considered as haplotype recipients only, whereas other populations or meta-populations were considered as both donors and recipients. Note that this differs from the ChromoPainter v.1 approach, in which all individuals were considered as donors and recipients of haplotypes at the same time, and only self-copying was forbidden.

Painting samples for the target population and ‘copy vectors’ for other (meta) populations called ‘surrogates’ served as an input of GLOBETROTTER, which was run according to section 6 of the instruction manual. The following settings were used: no standardizing by a ‘NULL’ individual (null.ind.0); five iterations of admixture date and proportion/source estimation (num.mixing.iterations.5); at each iteration, any surrogates that contributed  $\leq 0.1\%$  to the target population were removed (props.cutoff.0.001); the  $x$  axis of co-ancestry curves spanned the range from 0–50 cM (curve.range.1.50), with bins of 0.1 cM (bin.width.0.1). Confidence intervals (95%) for admixture dates were calculated based on 100 bootstrap replicates. Alternatively, when using separate populations as haplotype donors, the setting ‘standardizing by a ‘NULL’ individual’ was turned on to take account of potential bottleneck effects. A generation time of 29 years was used in all dating calculations<sup>5,29</sup>.

The GLOBETROTTER software is able to date no more than two admixture events<sup>25</sup>, and we therefore had to reduce the complexity of original Na-Dene-speaking populations that probably experienced more than two major waves of admixture. For that purpose, only a subset of Na-Dene-speaking individuals was used for the GLOBETROTTER analysis: those with prior evidence of



elevated Palaeo-Eskimo ancestry (Supplementary Information section 6) and with no more than 10% West Eurasian ancestry estimated with ADMIXTURE (Extended Data Fig. 8). We also performed a similar analysis with ALDER (Supplementary Information section 12).

**Rare-allele-sharing statistics.** To quantify rare-allele sharing, we developed RASS. Essentially, RASS is similar to the outgroup  $f_3$  statistic, but it is ascertained on rare 'non-outgroup' alleles in a set of reference populations. Specifically, we define

$$\text{RASS}(x, y; \{\text{references}, \text{outgroup}\}) = \frac{1}{L} \sum_i x_i y_i$$

in which the sum runs over all sites with derived allele count below some cutoff (say five or less) within the reference and outgroup populations,  $x_i$  is the derived allele frequency in the test individual,  $y_i$  is the derived allele frequency in the reference population, and  $L$  is the number of sites in the sum (excluding missing data). Here, the outgroup (the African meta-population) is used to polarize derived versus ancestral alleles. We look at the outgroup population, and take the majority allele in that outgroup population to specify which should be the majority allele for the ascertainment. If the majority of outgroup chromosomes have the non-reference allele, then the ascertainment is done on the reference allele being rare (instead of the non-reference allele). Standard errors are computed using a chromosome-wise weighted block jackknife. See Supplementary Information section 8 for details. We note that this method, in contrast to PCA, is not affected by genetic drift within the test individuals, because the ascertainment of allele frequency is carried out only in the reference populations. Source code for the programs used to perform rare-allele-sharing analysis is available at <https://github.com/TCLamnidis/RASools> and <https://github.com/stschiff/rarecoal-tools>.

**Demographic modelling.** We used the qpGraph method<sup>8</sup> to explore models that are consistent with  $f$ -statistics. We started by using qpGraph v.5052 to build a backbone graph of eight populations that represent several major branches of human ancestry (African, European, Southeast Asian, Siberian, Chukotko-Kamchatkan, Eskimo–Aleut, Athabaskan, and First Peoples; Supplementary Information section 10). One difficulty in estimating admixture graphs for closely related populations, such as the ones studied here, is the fact that typically many different graphs fit the data equally well. We therefore used an iterative approach in which we kept not only the best-fitting model at each stage in the model development, but also all fitting models that connected a given set of populations. We then used this backbone to map several ancient populations onto the graph, and, in particular, we varied all possible topologies of the subgraph that connects Chukotko-Kamchatkan, Saqqaq, ancient Eskimo–Aleut, and ancient Athabaskan populations. With 224 models tested (varying both the Neo-Eskimo population as well as the PPE topology), we found that the best-fitting topology of this PPE clade had Chukotko-Kamchatkan speakers splitting off first, then the PPE-admixture source in Athabaskans, then the ancient Saqqaq and the PPE-source in ancient Eskimo–Aleuts: (C-K, (ATH<sub>PPE</sub>, (SAQ, E-APPE))); see Supplementary Information section 10). We further confirmed these models by testing 133,380 models derived from the main model, but replacing the meta-populations with concrete populations (see Supplementary Information section 10).

We used a newly developed version of the Rarecoal program<sup>9</sup> (<https://github.com/stschiff/rarecoal>) to derive a timed admixture graph for meta-populations (Fig. 2 and Supplementary Information section 9). We started with a simple graph connecting Europeans, Southeast Asians, and Southern First Peoples, and inferred maximum-likelihood branch population sizes and split times. We then iteratively added Core Siberians (indigenous Siberians excluding Chukotkan and Kamchatkan groups), and Chukotko-Kamchatkan, Northern First Peoples, Aleut, Yup'ik/Inuit, and Northern Athabaskan groups. After each addition, we re-optimized the tree and inspected the fits of the model to the data. When we

observed a notable deviation between model and data for a particular pairwise allele-sharing probability, we added admixture edges (Supplementary Information section 9), which were in all cases consistent with the final qpGraph model graph. We then tested several positions for the Saqqaq genome to merge onto the tree, and found that the maximum-likelihood position was one in which Saqqaq merges on the common ancestor of Eskimo–Aleut branches, before interactions with Northern Peoples but after the gene flow from that same lineage into Athabaskans (see Fig. 2b). We also derived confidence intervals and we corrected likelihood model comparisons using a correction for genetic linkage correlations in the data, using a jackknife procedure, as described in Supplementary Information section 9. We then also mapped the ancient Aleut and ancient Athabaskan individuals onto the tree.

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

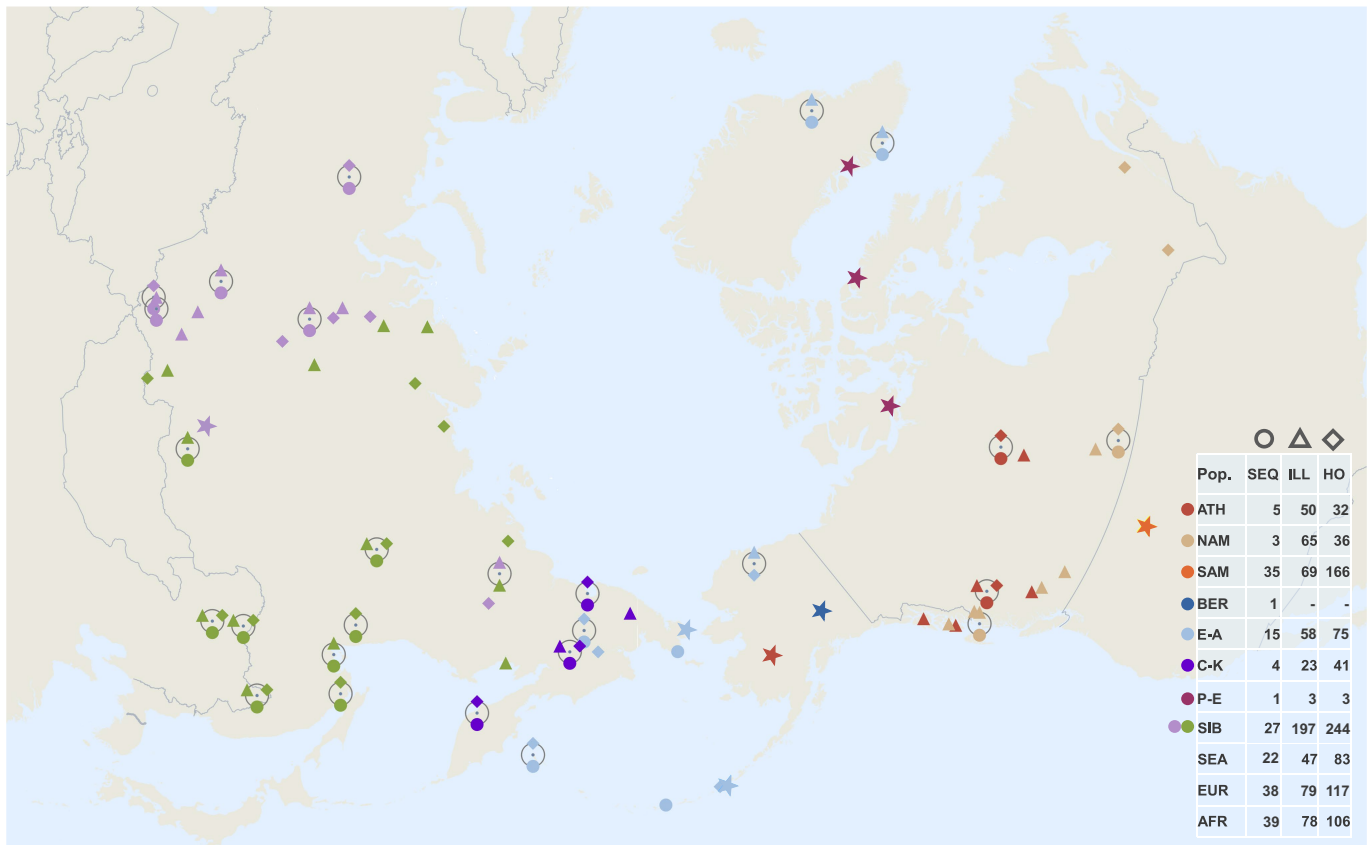
## Data availability

Raw sequence data (.bam files) from the 48 ancient individuals that we studied here are available from the European Nucleotide Archive under accession number PRJEB30575. The genotype data for the Iñupiat were obtained through informed consent, which does not allow us to provide the data through public or controlled-access data repositories; it also does not allow analyses of phenotypic traits, or commercial use of the data. To protect the privacy of participants and ensure that their wishes with respect to data usage are followed, researchers wishing to use data from the Iñupiat samples should contact M.G.H. ([ghayes@northwestern.edu](mailto:ghayes@northwestern.edu)) and D.A.B. ([deborah.bolnick@uconn.edu](mailto:deborah.bolnick@uconn.edu)), who can then arrange to share the data with researchers who can affirm that they will abide by the relevant conditions through a signed data-sharing agreement. The SNP genotyping data for West Siberians (Enets, Kets, Nganasans, and Selkups) are publicly available at the Edmond database, under the permalink <https://doi.org/10.17617/3.1z>.

## Code availability

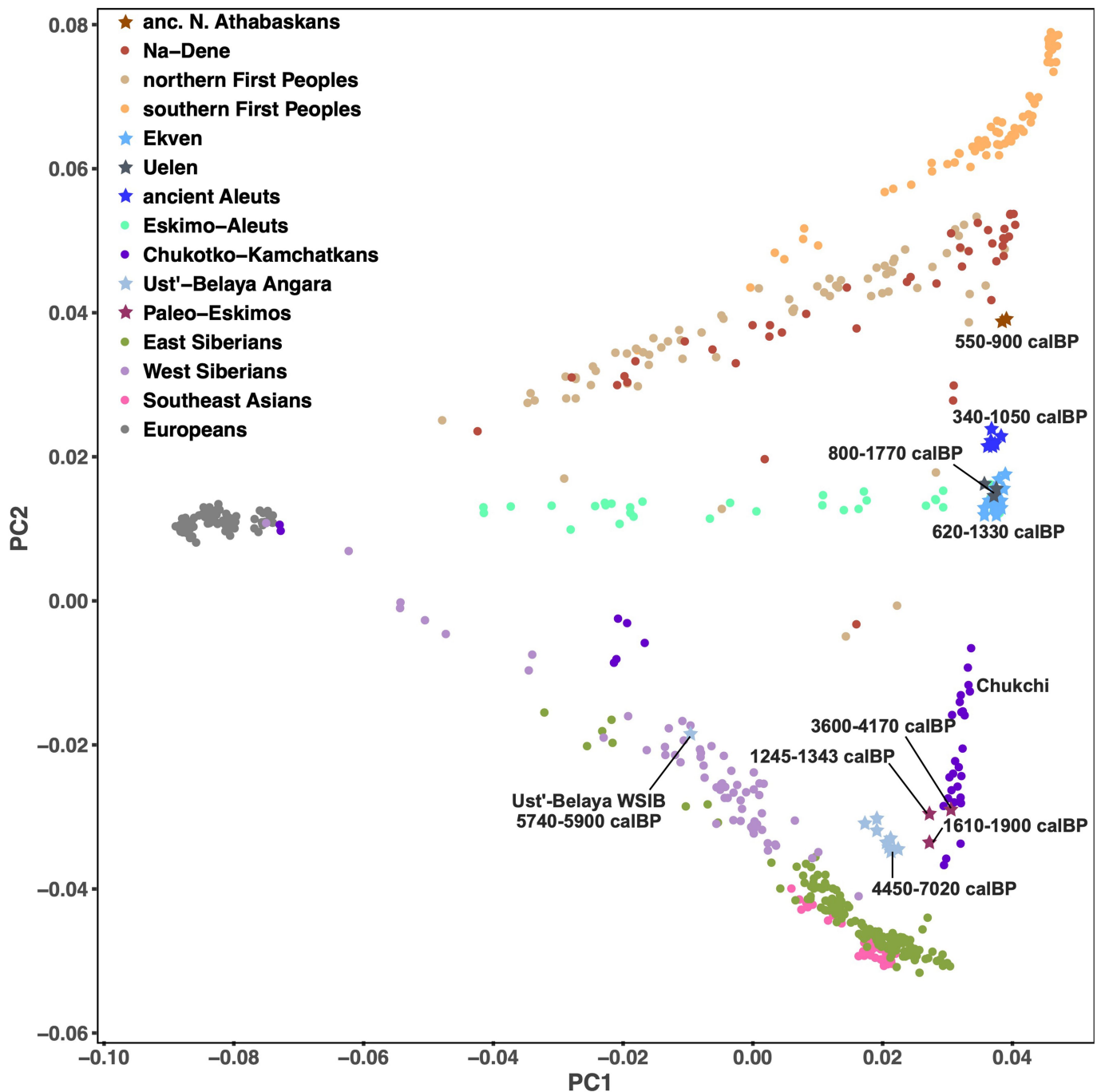
Custom code used in this manuscript is available at dedicated GitHub repositories: rarecoal (<https://github.com/stschiff/rarecoal>), rarecoal-tools (<https://github.com/stschiff/rarecoal-tools>), and RAS-tools (<https://github.com/TCLamnidis/RASools>).

- Flegontov, P. et al. Genomic study of the Ket: a Paleo-Eskimo-related ethnic group with significant ancient North Eurasian ancestry. *Sci. Rep.* **6**, 20768 (2016).
- Mallick, S. et al. The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. *Nature* **538**, 201–206 (2016).
- Li, H. & Durbin, R. Inference of human population history from individual whole-genome sequences. *Nature* **475**, 493–496 (2011).
- Rasmussen, M. et al. The genome of a Late Pleistocene human from a Clovis burial site in western Montana. *Nature* **506**, 225–229 (2014).
- Raghavan, M. et al. Upper Palaeolithic Siberian genome reveals dual ancestry of Native Americans. *Nature* **505**, 87–91 (2014).
- Lazaridis, I. et al. Ancient human genomes suggest three ancestral populations for present-day Europeans. *Nature* **513**, 409–413 (2014).
- O'Connell, J. et al. A general approach for haplotype phasing across the full spectrum of relatedness. *PLoS Genet.* **10**, e1004234 (2014).
- Purcell, S. et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
- Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* **19**, 1655–1664 (2009).
- Loh, P. R. et al. Inferring admixture histories of human populations using linkage disequilibrium. *Genetics* **193**, 1233–1254 (2013).
- Lazaridis, I. et al. Genomic insights into the origin of farming in the ancient Near East. *Nature* **536**, 419–424 (2016).
- Verdu, P. et al. Patterns of admixture and population structure in native populations of Northwest North America. *PLoS Genet.* **10**, e1004530 (2014).



**Extended Data Fig. 1 | Geographic locations of Siberian and North American populations used in this study.** The three main datasets are as follows (Supplementary Tables 4, 5): (1) a set based on the Affymetrix Human Origins genotyping array, including alternatively pseudo-haploid or diploid genotypes for the ancient Saqqaq individual<sup>1</sup>; diploid genotypes for the ancient Clovis<sup>35</sup> individual, together with 1240K SNP capture pseudo-haploid data from 6 ancient Aleuts who had the highest coverage; 2 unrelated ancient Athabaskans; 19 ancient Old Bering Sea individuals from the Ekven and Uelen sites; the Middle Dorset and Late Dorset Palaeo-Eskimo individuals; and the ancient Ust'-Belaya Angara population of 9 individuals (Supplementary Table 1); (2) a set based on various Illumina arrays, including Saqqaq and the other ancient samples; and (3) a whole-genome dataset of 190 individuals from 87 populations, including the Saqqaq individual, 1 ancient Athabaskan individual (I5319), and 1 ancient Aleut individual (I0719), for whom we generated complete

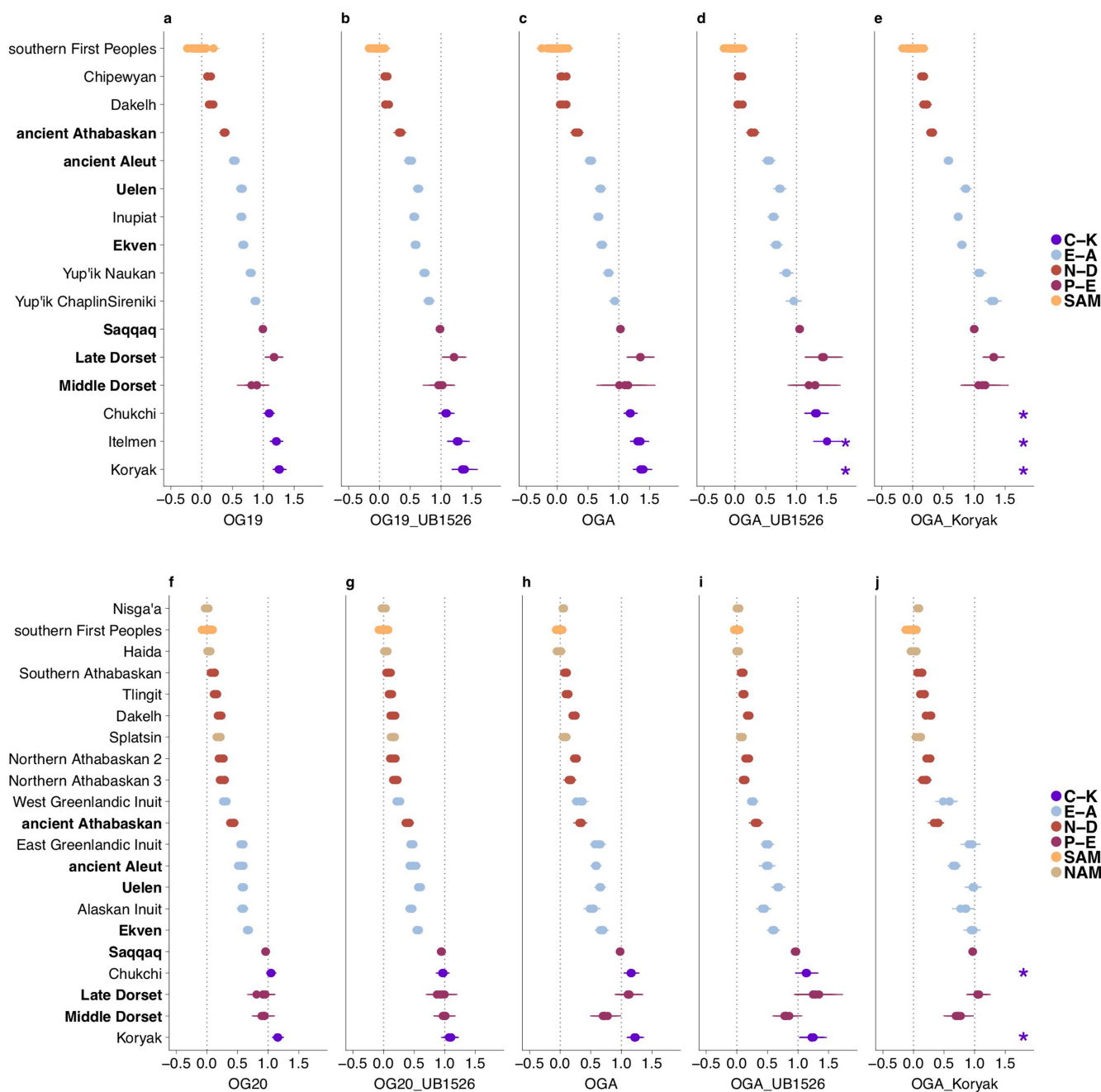
genomes with  $6.1\times$  and  $2.3\times$  coverage, respectively (Supplementary Table 1). The dataset composition—that is, the number of individuals in each meta-population—is shown in the table on the right. Locations of individuals with whole-genome sequencing data (SEQ) are shown with circles, and those of Illumina (ILL) and HumanOrigins (HO) SNP-array samples with triangles and diamonds, respectively. Meta-populations are colour-coded in a similar way throughout all figures and designated as follows: Na-Dene speakers (abbreviated as ATH), other northern Native Americans (NAM) (alternatively known as Northern First Peoples), Southern First Peoples (SAM), Ancient Beringians (BER), Eskimo-Aleut speakers (E-A), Chukotko-Kamchatkan speakers (C-K), Palaeo-Eskimos (P-E), West and East Siberians (WSIB and ESIB), Southeast Asians (SEA), Europeans (EUR), and Africans (AFR). The locations of the Saqqaq, Dorset, and other ancient individuals are shown as stars that are coloured to reflect their meta-population affiliation.



**Extended Data Fig. 2 | PCA based on the Illumina dataset.** A plot of two principal components (PC1 versus PC2) calculated using PLINK2 is shown (linkage disequilibrium pruning was not applied). No outliers were excluded for this analysis, which was based on 642 individuals and 524,830 loci. The following meta-populations most relevant for our study are plotted: present-day Eskimo-Aleut and Chukotko-Kamchatkan speakers, ancient Chukotkan Neo-Eskimos (Ekven and Uelen sites), ancient Aleuts, Palaeo-Eskimos (the Saqqaq, Middle Dorset, and Late

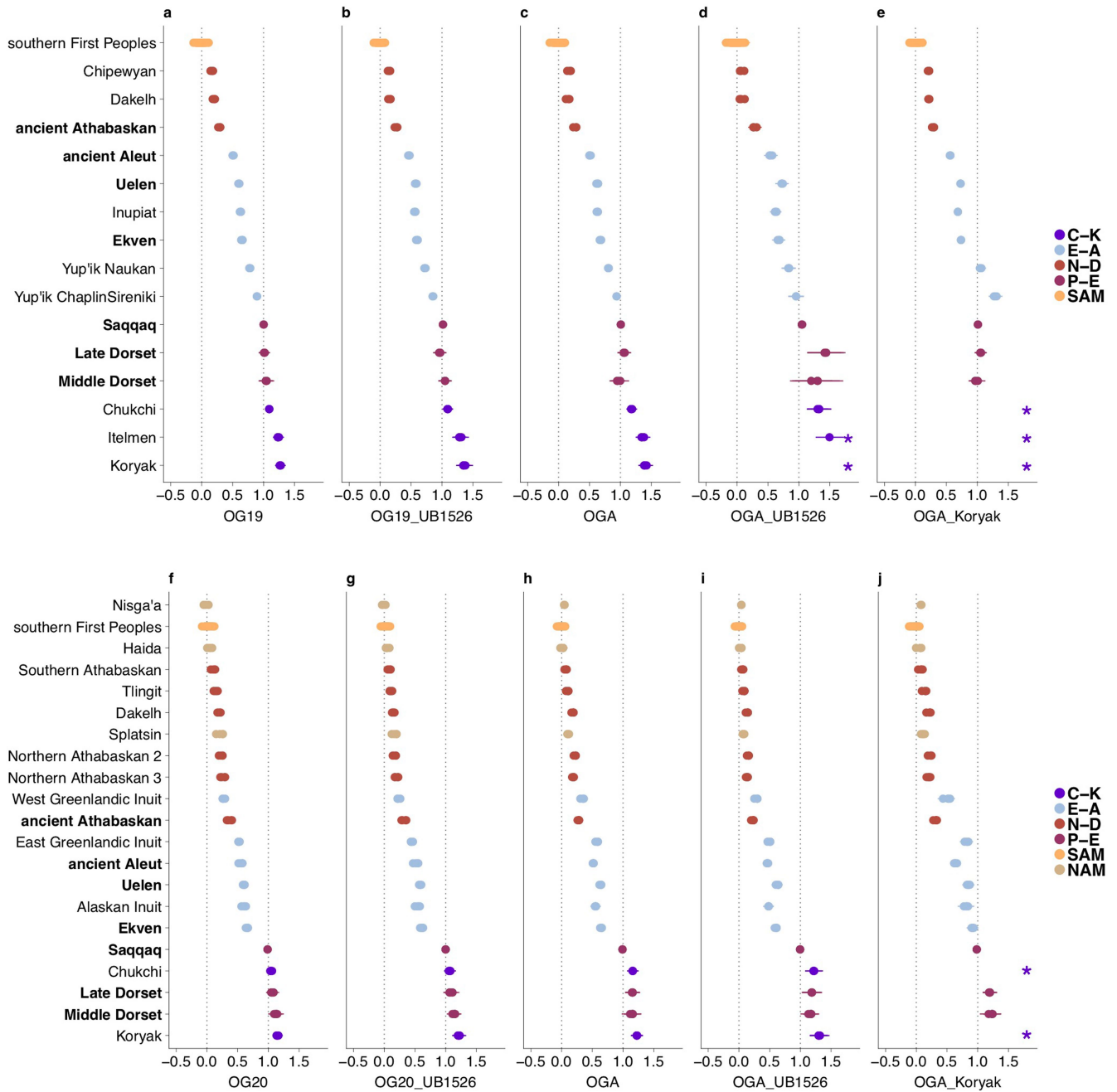
Dorset individuals), ancient Northern Athabaskans, present-day Na-Dene speakers, Northern and Southern First Peoples, West and East Siberians, the Ust'-Belaya Angara ancient Siberian group, Southeast Asians, and Europeans. Radiocarbon dates in cal. yr BP are shown for ancient samples. For individuals, 95% confidence intervals are shown, and for groups of individuals, minimal and maximal median dates among individuals are shown.





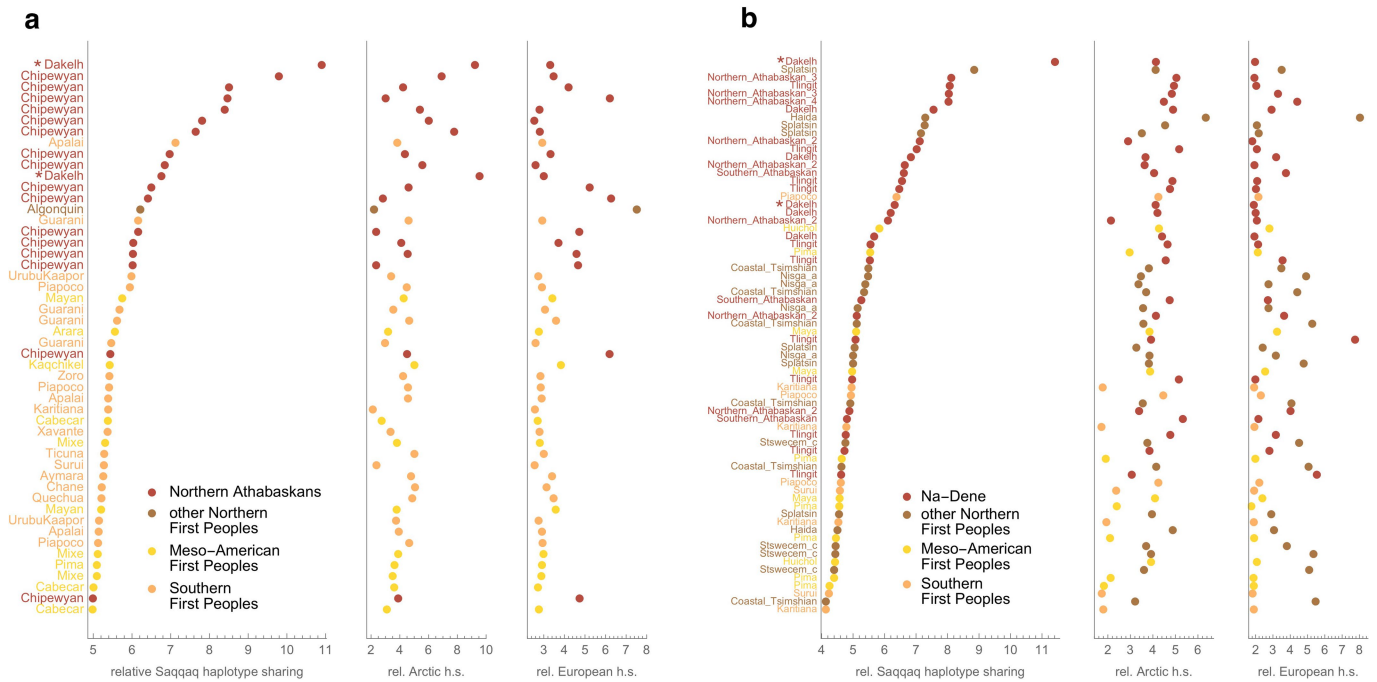
**Extended Data Fig. 3 | Ancestry proportions in American, Chukotkan, and Kamchatkan populations. a–j,** The HumanOrigins (a–e) and Illumina (f–j) datasets without transition polymorphisms are shown. Five alternative outgroup sets are indicated below the plots and described in detail in the Methods and Supplementary Information section 5. Bold formatting denotes ancient target groups. Saqqaq (pseudo-haploid genotype calls) was considered as a Palaeo-Eskimo source for all populations apart from Saqqaq itself (for which Late Dorset was used as a source) and alternative First Peoples sources were as follows: Mixe, Guarani, or Karitiana for the HumanOrigins dataset; Nisga’a, Mixtec,

Pima, or Karitiana for the Illumina dataset. To visualize both systematic and statistical errors, ancestry proportions inferred by qpAdm and their standard errors are shown for all triplets including these different First Peoples sources, or for many alternative target groups in the case of Southern First Peoples (single standard error intervals are plotted here). Asterisks indicate ancestry proportions greater than 150% (inappropriate models). Meta-populations are colour-coded according to the legend and abbreviated as before (N-D, Na-Dene speakers). Target group sizes in the HumanOrigins dataset ranged from 1 to 23 individuals (average 5.6), and in the Illumina dataset they ranged from 1 to 16 individuals (average 5.1).



**Extended Data Fig. 4 | Ancestry proportions in American, Chukotkan, and Kamchatkan populations. a–j,** Similar analysis as in Extended Data Fig. 3, but including transition polymorphisms. Target population

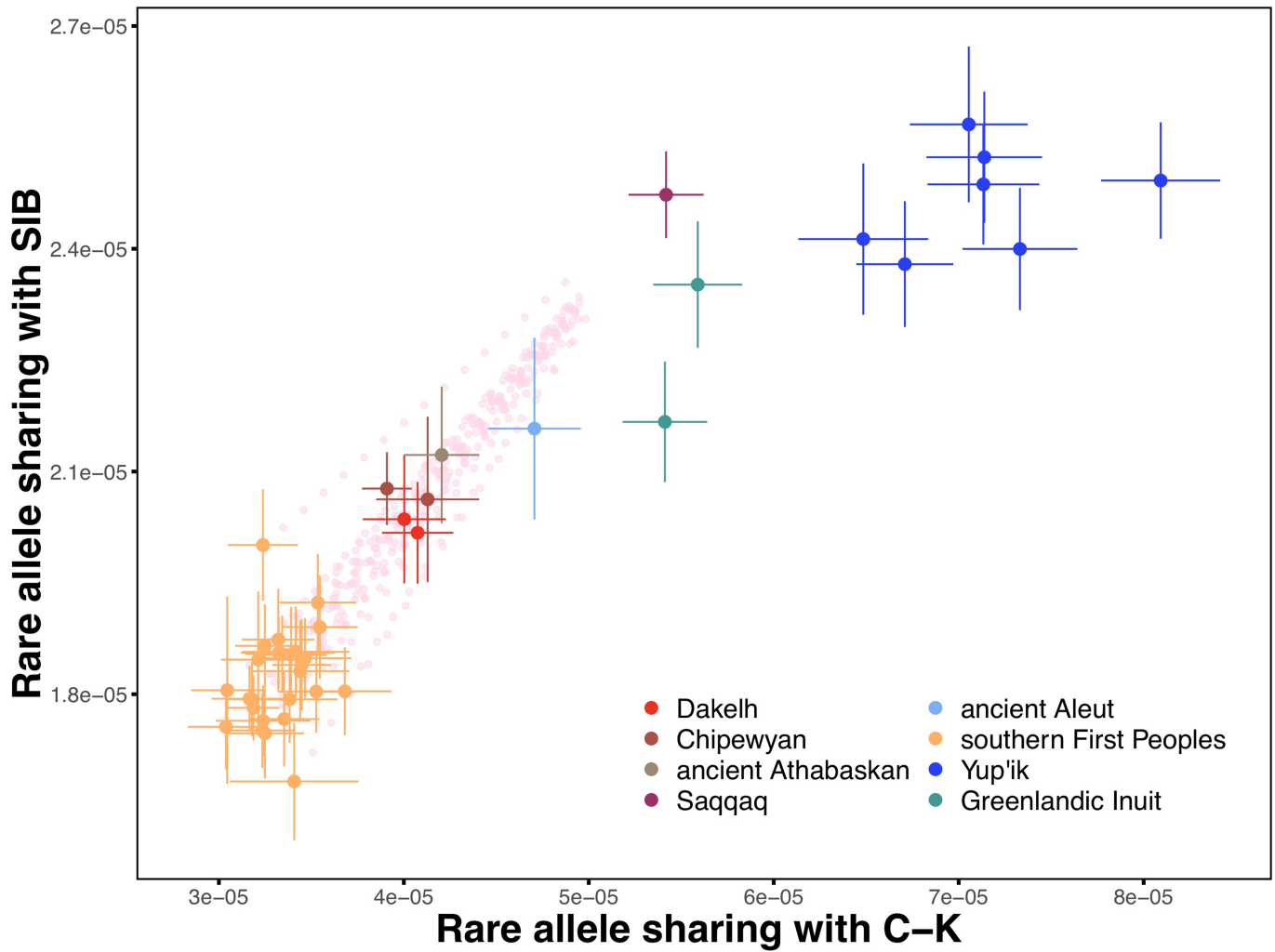
sizes in the HumanOrigins dataset (a–e) ranged from 1–23 individuals (average 5.6), and in the Illumina dataset (f–j) they ranged from 1–16 individuals (average 5.1).



**Extended Data Fig. 5 | Relative Saqqaq, Arctic, and European haplotype-sharing statistics for American individuals.** **a, b,** Results are shown for the Human Origins (**a**) and Illumina (**b**) datasets, normalized using the African meta-population. Both Eskimo–Aleut- and Chukotko–Kamchatkan-speaking groups contributed to the Arctic HSS. The same statistics and statistics with other normalizers are shown in the form of two-dimensional plots in Supplementary Information section 6. Two Dakelh (Northern Athabaskan) individuals with whole-genome sequencing data<sup>5</sup> were included in both datasets and are marked with asterisks. The plots based on both datasets demonstrate that Na-Dene speakers have the highest relative Saqqaq HSS. One Haida and three Spltasin individuals also demonstrate outlying Saqqaq HSSs (**b**); however,

these individuals contrast with a majority of non-Na-Dene-speaking Northern First Peoples, and Palaeo-Eskimo ancestry in these individuals may be explained by recent interaction with Na-Dene speakers living in close proximity<sup>43</sup>. The Haida outlier demonstrates the maximal Arctic HSS among all First Peoples, and their Arctic ancestry has contributed to their elevated Saqqaq HSS. Saqqaq, Arctic, and European statistics are largely uncorrelated in First Peoples: Pearson’s correlation coefficients for Saqqaq versus Arctic relative HSSs are 0.56 among all First Peoples and 0.64 among Northern First Peoples in the case of the Illumina dataset, and 0.66 and 0.72, respectively, in the case of the HumanOrigins dataset. h.s., haplotype sharing.

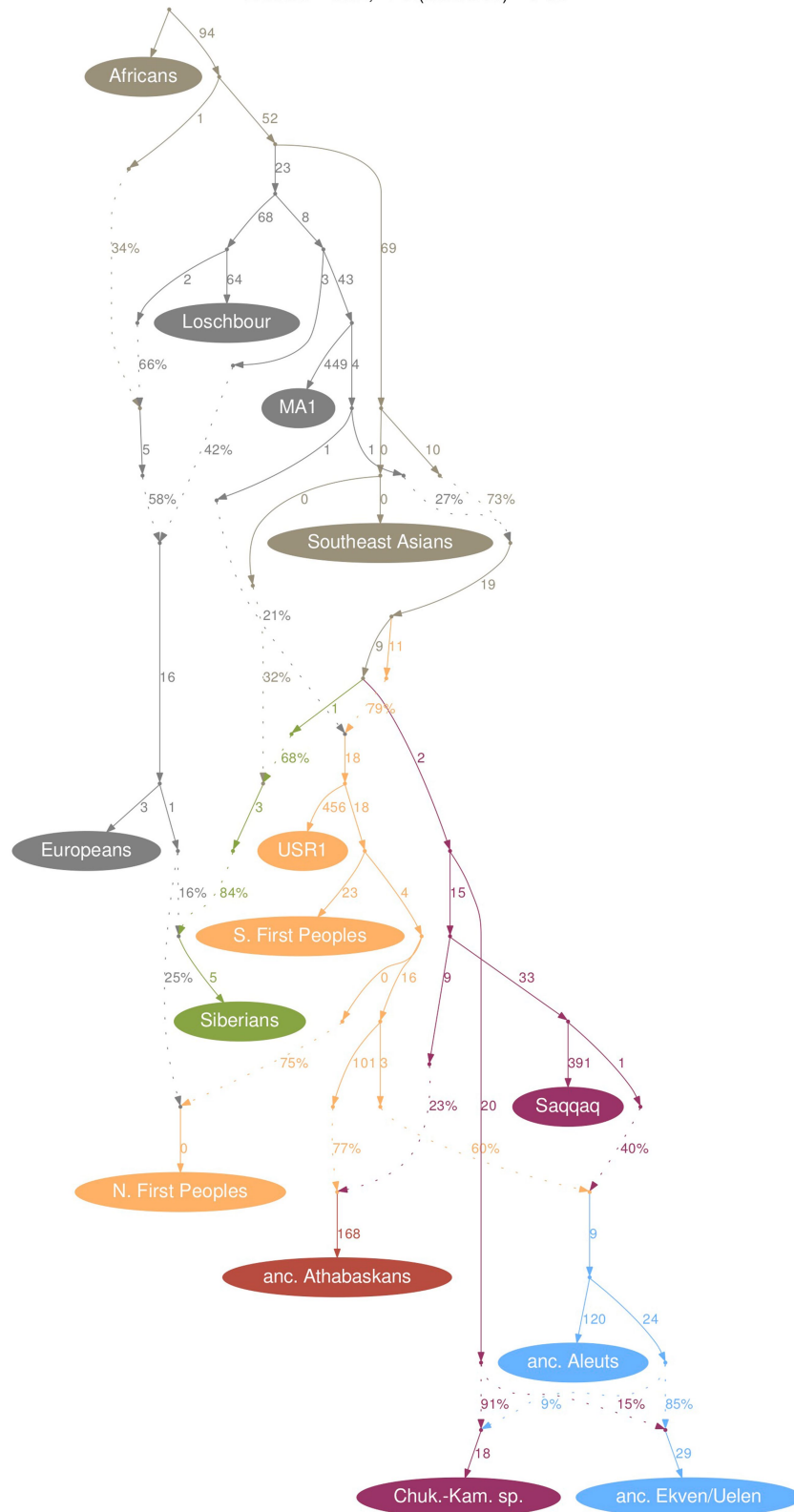




**Extended Data Fig. 6 | Rare-allele-sharing analysis.** A two-dimensional plot of Chukotko-Kamchatkan and Siberian rare-allele-sharing statistics for First Peoples, Na-Dene-speaking, Eskimo-Aleut-speaking, and Palaeo-Eskimo individuals. Rare alleles occurring from 2 to 5 times in the reference set of 238 haploid genomes (0.8–2.1% frequency) contributed to the statistics; the Chukchi individual was dropped from the Chukotko-Kamchatkan reference group, and the transversion-only dataset was used. Thus, this analysis was based on 918,474 loci. The sample size for this analysis equals  $238 + 2$  haploid genomes in a target individual, as

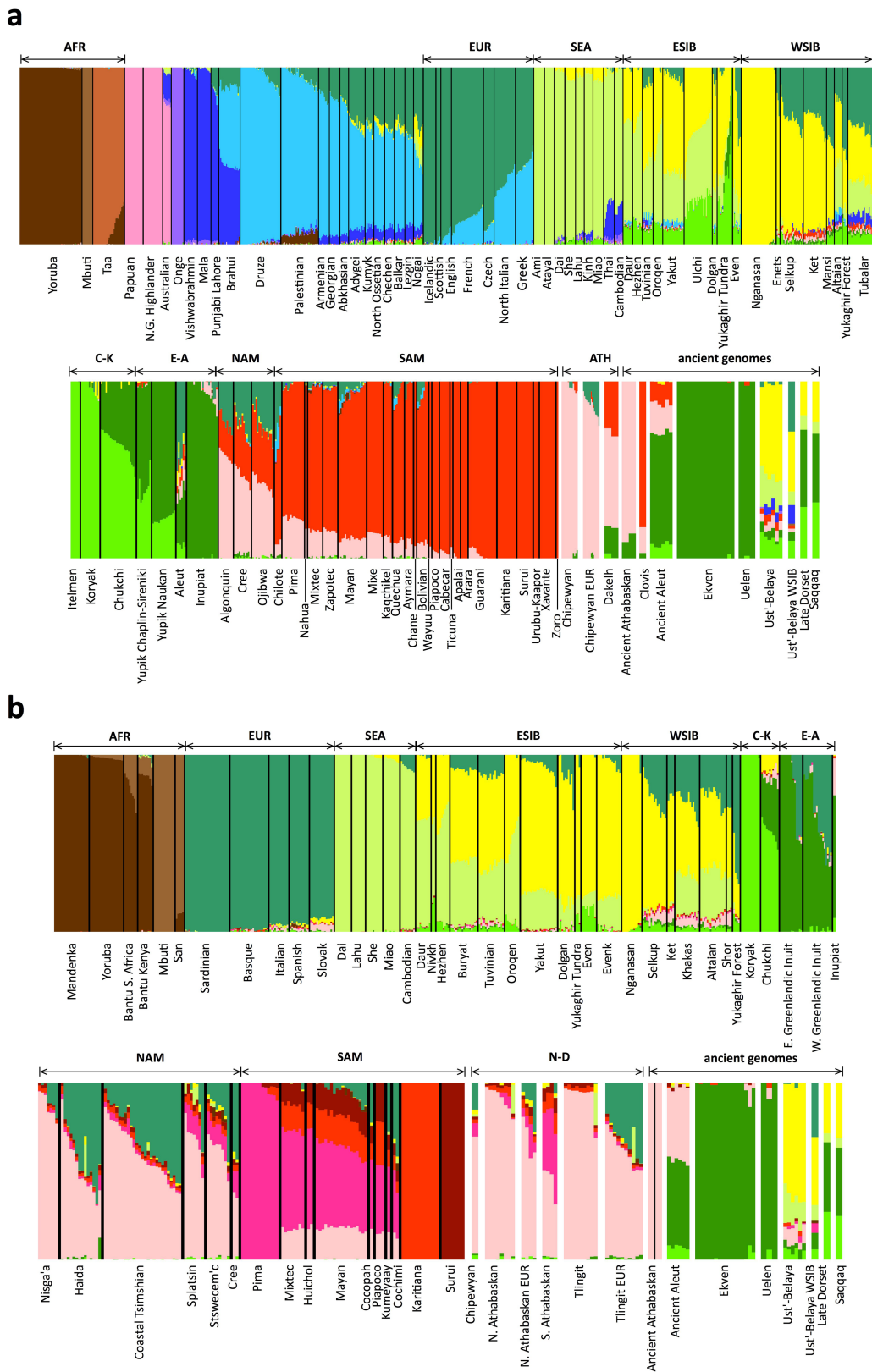
individuals were analysed separately. Standard deviations were calculated using a jackknife approach, with chromosomes used as resampling blocks. Single standard error intervals and means are plotted. Populations and meta-populations are colour-coded according to the legend. Rare-allele-sharing statistics for simulated mixtures of any present-day southern Native American individual and the Saqqaq individual (from 5–75% Saqqaq ancestry, with 5% increments) are plotted as semi-transparent pink circles. Plots for the 2–10 allele frequency range and other versions are shown in Supplementary Information section 8.

Z-score = 3.12,  $-1 \cdot \ln(\text{likelihood}) = 74.8$



**Extended Data Fig. 7 | An admixture graph connecting various modern meta-populations and ancient populations or individuals.** The graph (see Supplementary Information section 10) features a simplified three-component model for Europeans as previously suggested<sup>37</sup>, and two gene flows from a European lineage related to the ancient Siberian genome

MA-1<sup>36</sup> into Native Americans and Siberians. The topology within the PPE clade was obtained by cycling through dozens of trees with all possible topologies of branches and admixture edges, and selecting the one with the highest support and no zero-length edges within the PPE clade.



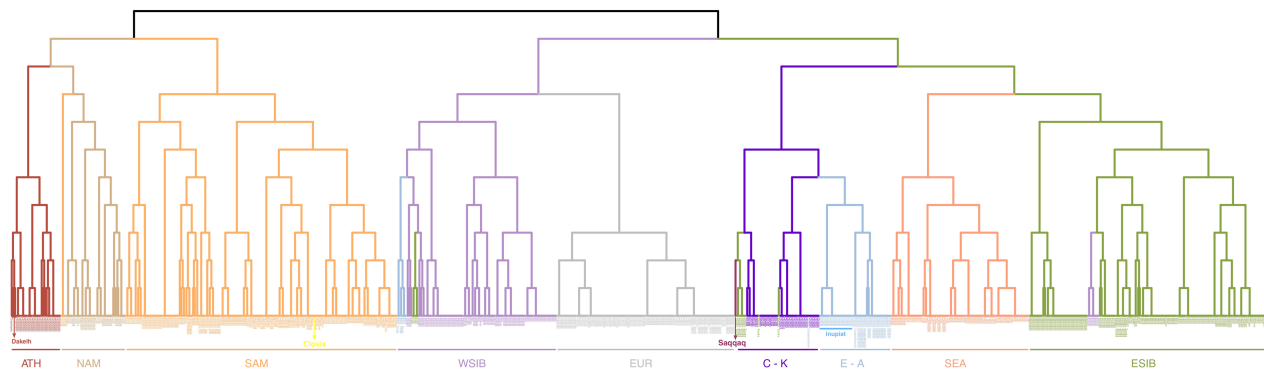
Extended Data Fig. 8 | See next page for caption.



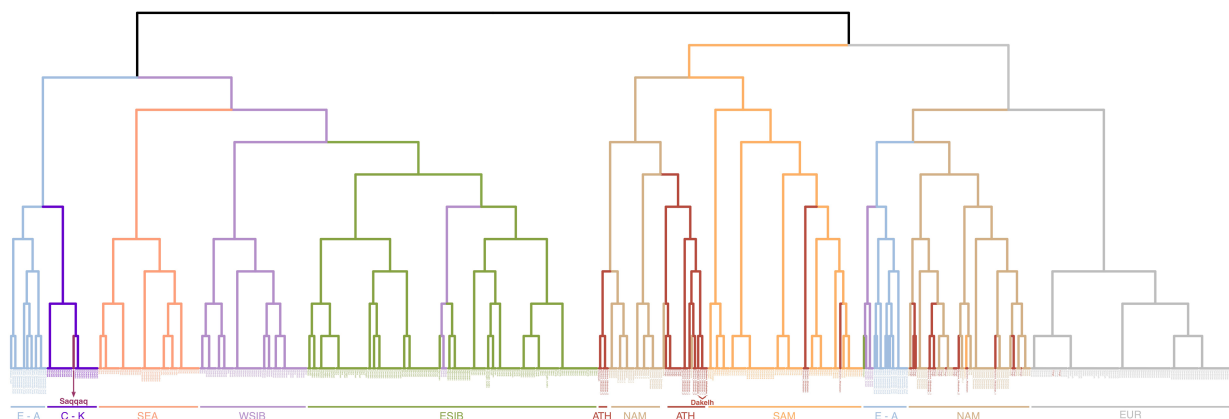
**Extended Data Fig. 8 | ADMIXTURE analysis. a, b,** Results are shown for the HumanOrigins (a) and Illumina (b) SNP-array datasets. The number of source populations in ADMIXTURE is 14 and 11 for the 2 datasets, respectively. One hundred iterations were calculated for each value of  $K$  from 5–20 (in which  $K$  is the number of ancestral populations), and the optimal  $K$  values were selected based on 10-fold cross-validation. Contributions from hypothetical ancestral populations are colour-coded, and meta-populations used in this study are indicated above the plot (abbreviations as before). Chipewyan or Northern Athabaskan and Tlingit individuals with European admixture are plotted in separate bars, as are ancient individuals: Clovis, Northern Athabaskans, Aleuts, Chukotkan

Neo-Eskimos (Ekven and Uelen sites), Saqqaq and Late Dorset Palaeo-Eskimos, and a genetically heterogeneous Ust'-Belaya Angara Siberian population (Ust'-Belaya WSIB, an individual I7760 who has a West Siberian genetic profile according to PCA and this ADMIXTURE analysis; Ust'-Belaya, the remaining eight individuals from the Ust'-Belaya Angara site who have a distinct genetic profile according to our PCA analysis). Outliers, including individuals admixed with Europeans and East Asians, were not removed from Na-Dene-speaking populations in the Illumina dataset (b) to preserve their maximal diversity. Outliers were removed for the purpose of other analyses that rely on pre-defined populations (for example, qpAdm and  $f_4$ -statistics).

a



b



**Extended Data Fig. 9 | Clustering trees of individuals, computed by fineSTRUCTURE.** **a, b**, The trees are based on co-ancestry matrices of counts of shared haplotypes. Reduced versions of the HumanOrigins (**a**) and Illumina (**b**) SNP-array datasets were used (Supplementary Table 5), including only the following meta-populations that were most relevant for our study: Eskimo–Aleut speakers, Chukotko-Kamchatkan speakers, Na-Dene speakers, Northern First Peoples, Southern First Peoples, West Siberians, East Siberians, Southeast Asians, and Europeans. Meta-population affiliation is colour-coded for individuals. Inupiat individuals genotyped in this study are marked with a blue line. The two Dakelh (Northern Athabaskan) individuals with sequenced genomes are also indicated, as well as the ancient individuals—Clovis within the Southern

First Peoples clade and Saqqaq within the Chukotko-Kamchatkan clade. Most members of each clade belong to the meta-populations indicated, with a few exceptions. First (**a**), Altaians fall into the ESIB clade, some Chilote fall into the NAM, and Aleuts fall into the WSIB clades (the two latter cases might be explained by extensive European ancestry in Chilote and in Aleuts (Extended Data Fig. 8a), which drives this clustering). Second (**b**), some Selkups fall into the ESIB clade, all four Southern Athabaskan speakers cluster with South Americans (reflecting their substantial South American ancestry (Extended Data Fig. 8b)), one Haida individual clusters with Na-Dene speakers, and five Northern Athabaskan speakers cluster with other Northern First Peoples.