SUPPLEMENTARY INFORMATION

doi:10.1038/nature13025

Table of Contents

1. Archaeology of the Anzick Site, Montana

The Anzick site, Montana (24PA506) is located one mile south of Wilsall, Park County, Montana in an intermontane basin just east of the front of the Rocky Mountains[1-5]. Specifically, the Anzick site is situated along Flathead Creek just above its confluence with the Shields River. Here, fragmentary human remains and artefacts were found in sediments that lay adjacent to a north-facing escarpment composed of Cretaceous conglomerates, sandstones, and siltstones. Specifically, two fragmentary sets of human remains were uncovered: 1) an ochre-stained cranium and other elements of a 1-2 year-old child found in association with 115 red ochre-covered Clovis artefacts, and 2) unstained cranial fragments of a 7-8 year-old child found in a different location, approximately 6m east of the skeletal remains and not associated with the Clovis artefacts[4].

For a detailed history of the investigation of the site see Lahren[2] and Owsley and Hunt[4]. Briefly, the Clovis artefacts and human remains were approximately 2.5m below the surface of the talus in a gray, fine-grained stratum approximately 50cm thick. The bifaces and projectile points were found tightly stacked within a pit, about 1m in diameter, that had been dug into the gray fine-grained sediments. The artefacts were heavily coated with red ochre. Ochre-covered human bones were found beneath the artefacts.

The Clovis lithic artefacts were recovered and included 84 complete and fragmentary bifacially flaked cores and projectile point preforms, eight fluted projectile points, six unifacial cutting and scraping tools, and two pieces of shatter from biface reduction[6,7]. These artefacts were made on stone obtained from at least six different locations that are within about 200km of the site[6]. Associated with the human skeletal remains and flaked stone artefacts were 15 total osseous fragments. Two complete shafts (one of which consisted of 2 mending fragments and the other consisted of 4 mending fragments, accounting for 6 of the 15) and additionally there were 4 beveled ends and 5 midsection fragments[6,8]. The stone and osseous artefacts are technologically Clovis and constitute a functional tool kit; some of the stone artefacts show evidence of use and wear, and several of the stone and osseous tools may have been intentionally broken at the time of interment[6,8].

The ochre-stained human remains (Anzick-1) consisted of 28 cranial fragments that form all portions of the vault (dolichocranic), the left clavicle, left fourth rib, and right third and fourth ribs[4]. This individual at the time of death was between one and two years old. The bones are unburned and represent an inhumation.

Even though the cranial elements were directly dated and the technology of the stone and osseous tools is clearly Clovis, two of the osseous tools were radiocarbon dated (Figure 1 main text). Morrow and Fiedel[6] obtained two radiocarbon dates on collagen extracted from two osseous tools, and yielded an average age of 11,040±35 ^{14}C years BP (see Extended Data Table 1 for all dates presented in in the following). To test the ages obtained by Morrow and Fiedel[6], we ^{14}C dated XAD-purified collagen from one of the osseous artefacts and obtained a date of 11,025±30 ^{14}C years BP (UCIAMS-61661). These three ^{14}C measurements on osseous artefacts are statistically identical and yield an average ^{14}C age of 11,035±45 ^{14}C years, which calibrates as 13,039 to 12,763 calendar years BP. The date on Anzick-1 cranial bone (10,705±35) differs from the average date for the rods (11,035±45). Their respective calibrated age ranges are 12,722 to 12,590 calendar years BP (Anzick-1) and 13,039 to 12,763 calendar years BP (antler rods) — values that do not overlap at the 95.4% confidence interval. It has been postulated that this age discrepancy may indicate 1) the osseous artefacts are heirlooms or artefacts of ritual significance passed down through generations[9] 2) contamination in the bone or 3) geologically ancient bone being used several decades later than the antler's biological formation. The ages of both the ochre-stained cranial fragments and the osseous tools are within the accepted age range of Clovis[3]. Based on morphological characteristics, Morrow and Fiedel[6] suggested that the osseous tools were made of elk antler. The mitochondrial DNA (mtDNA) control region of one of the osseous tools was sequenced and found to be from elk (*Cervus elaphus*) belonging to haplogroup MM004 (SI section 3.4). This haplogroup was common in Beringia (north-eastern Siberia and north-western North America) in the late Pleistocene, while its distribution south of the ice sheet remains unknown. This osseous tool represents the oldest appearance of elk south of the continental ice sheets, at 11,000 ^{14}C years BP.

The unstained human remains (Anzick-2) consist of four articulating pieces of the posterior left and right parietals and the occipital squamous from a juvenile that was 6 to 8 years old at the time of death[4]. These remains were found on the modern ground surface about 6m east of the Anzick-1 skeleton. The absence of red-ochre staining on the bones shows that they are not associated with the ochre-stained human remains (Anzick-1) and Clovis artefacts. This was confirmed by radiocarbon dating, with the unstained elements dating to 8610±40 ^{14}C years BP (average of five ages) or 9530 to 9600 calendar years BP. and the stained elements dating to 10,710±35 ^{14}C years BP (CAMS-80538) or 12,581 to 12,656 calendar years BP.

2. Dating

2.1 Pretreatment of Bone Collagen for AMS ^{14}C Dating and Stable Isotope Analysis

Bone samples were broken into approximately 4-5 mm fragments and decalcified in 4°C 0.5*N* HCl over 3 to 5 days; after washing in deionized water, the decalcified collagen was extracted with 0.1% KOH at 4°C for 24-30 hours and washed to neutrality with DI water. The KOH-extracted, decalcified collagen's per cent pseudomorph was recorded, and the collagen freeze-dried to determine per cent yield of collagen relative to modern bone. Approximately 20-50 mg of decalcified, KOH-extracted collagen were heated at 90°C in 0.02*N* HCl to dissolve (gelatinize) the collagen. Heating continued only until the collagen dissolved, usually 5 to 30 minutes. After filtering the gelatin solution through a 0.45 μm Millex Durapore filter, the solution was freeze-dried and a per cent gelatinization and per cent weight yield were determined. Approximately 5-10 mg of gelatin were hydrolysed 22 hours at 110°C in distilled 6*N* HCl. The hydrolysate, containing free amino acids, fulvic acids, and insoluble inorganic and organic detritus, was passed through a 2 cm long X 5 mm diameter bed of XAD-2 resin in a solid phase extraction (SPE) column attached to a 0.45 μm Millex filter. The XAD column contained 100-200 μm diameter research grade XAD-2 from Serva Biochemicals (Cat. No. 42825). The bulk resin was initially wetted with acetone and washed voluminously with DI water and finally multiple washes with distilled 1*N* HCl. Individual SPE columns were packed with the XAD-2 as a slurry of resin and HCl. Each column was equilibrated with 50 ml of distilled 6*N* HCl and the washings discarded. The collagen hydrolysate as approximately 1 to 2 ml of HCl was pipetted onto the SPE XAD column and eluted into a glass tube. Following the initial sample aliquot, the column was washed with 5 ml of 6*N* HCl that was added to the original eluate. The XAD-purified collagen hydrolysate was dried by passing ultra high purity N_2 gas over the HCl solution, resulting in a viscous syrup. The dry hydrolysate was diluted with DI water and approximately 2 to 3 mg of amino acids were transferred to 6 mm OD X 4 mm ID X 20 mm long quartz tubes and dried under vacuum. Approximately 50 mg of purified CuO wire and 5-10 mg Ag were added to each quartz tube. Stock CuO wire (Fisher Scientific, Cat. No. C474-500) was first combusted in crucibles at 900°C and stored in Pyrex tubes that were subsequently combusted at 570°C immediately before each use. Aesar 99.9995%, 30-60 mesh silver powder (Cat. No. 11408) was used without additional purification. All glass pipettes, beakers, and tubes were combusted at 550°C for 30 minutes before use. After evacuation to < 20 millitorr by vacuum pumping

through a liquid nitrogen trap, the tubes were sealed with a H_2/O_2 torch. The quartz sample tubes were combusted at 820°C for 2 hours and cooled from 820°C to 150°C at 60°C per hour.

2.2 Graphitization of CO_2 and AMS ^{14}C measurements

Following purification of the combustion and phosphoric acid hydrolysis products to remove water and non-condensable gases, 0.5 to 1 milligram of carbon as CO_2 was converted into graphite by the Fe-H_2 method[10]. Contemporary ^{14}C standards used for normalization are Oxalic Acid-I. Known-age bones used for calibrations included VIRI-I whale bone (Consensus age = 8331±6 ^{14}C years BP)[11] and Dent Mammoth bone (10,950±30 ^{14}C years BP)[3]. Respective chemistry and combustion backgrounds and blanks were determined by using > 70 ka collagen isolated from fossil *Eschrichtius* (Gray Whale) bone[3,12] and Sigma Chemical Company alanine (Sigma A-7627). Graphite targets were prepared and analysed at the Keck Carbon Cycle AMS Facility, Earth System Science Department, University of California-Irvine.

2.3 Measurement of Amino Acid Compositions for Bone and Collagen

Quantitative amino acid analyses were made by Margaret Condron at the Biopolymer Laboratory, David Geffen School of Medicine at UCLA. Approximately 1 mg of sample was hydrolysed by using 6*N* vapor-phase HCl. The hydrolysate was derivatised and subsequently analysed by reverse phase (RP)-HPLC using a highly fluorescent amino-reactive probe on a Waters Alliance HPLC system. Detection limits were approximately 1 picomole. Analyses are reported as residues per thousand (R/1000) for each amino acid and as total nanomoles (nmol) of amino acids per mg.

All dates are summarised in Extended Data Table 1.

3. Preliminary DNA screening for mtDNA

3.1 DNA Extraction

DNA from the Anzick-1 individual was extracted from 20 to 100mg of the petrous portion of the temporal bone, alongside 4 presumed elk samples also from the Anzick site. Extractions were performed according to strict aDNA-specific requirements in a dedicated ancient DNA facility located at the Centre of Excellence in GeoGenetics, University of Copenhagen. Full body suits, face masks, gloves, and hats sterilized by exposure to UV lamps (254 nm) were worn throughout the entire extraction and PCR assembly processes, and the workspace

surfaces and laboratory equipment were irradiated with UV lamps for 30 minutes and cleaned with concentrated bleach solution (2%) prior to all experiments. Extraction controls (water blanks) were included throughout all processes to monitor for exogenous, modern DNA contamination.

To reduce surface DNA contamination, the bone fragment was incubated in 10 mL 0.1 M HCl (UV irradiated at 254 nm for 20 min) with gentle agitation for 5 minutes and then rinsed twice in UV-irradiated double-distilled DNAse-free Millipore water. Following the water rinses, the sample was washed for 2 minutes in 10 mL 95% ethanol and then dried for 1 hour at room temperature. Fine-grained powder was obtained using sterile drill bits, which were bleach-sterilized and UV-irradiated. The resulting powder was soaked in 1.0 mL 0.5% bleach solution for 15 minutes under constant rotation. The bleach was discarded and the sample was rinsed in UV-irradiated water for 3 minutes. The bleaching process was repeated two times. The powder was placed in Yang buffer[13] (1M urea and 0.45M EDTA, pH 8.0) containing 250 mg Proteinase K and rotated at 55^0C for 24 hours. Next, the powder was centrifuged at 2,000 rpm for 5 minutes and the supernatant was concentrated in Amicon filters at 4,000 rpm for 5 minutes. To the recovered volume, 1,250 μL of Qiagen PB buffer was added and the DNAs were purified using QIAquick PCR purification kit following the manufacturer's protocol (Qiagen, Valencia, CA). Samples were eluted in 90 μL EB buffer.

3.2 Mitochondrial DNA Typing and Sequencing of Anzick-1 individual

DNA extracts were used as template as a preliminary screen for polymorphisms that define the five distinct founding mitochondrial DNA haplogroups A, B, C, D, and X[14,15]. For each PCR, 2 μL DNA extract (non-quantified) was used as template in a 25 μL reaction mixture containing 0.2 mM each dNTP, 0.3 μM each primer, 2.5 mM $MgSO_4$ and 2.0 U of HiFi Taq DNA polymerase (Life Sciences, Carlsbad, CA). Cycling conditions were: initial denaturing at 94^0C for 4 minutes, 45 cycles 94^0C for 15 seconds, 56^0C for 20 sec, and 68^0C for 30 sec, and a final extension at 72^0C for 7 min. The mtDNA sequence from hypervariable region I (HVR_1) was amplified using primers for six overlapping fragments, as described in Malmström *et al.*[16]. For HVR_1 fragments, PCRs were assembled in 25 μL volumes containing 2 μL DNA extract (non-quantified), 0.2 mM each dNTP, 0.3 μM each primer, 2.5 mM $MgCl_2$ and 2.0 U of SMART-Taq DNA polymerase (Naxo). The amplifications were carried out with an initial denaturation at 94^0C for 15 minutes, followed by 45 cycles of 94^0C for 30 sec, 54-60^0C for 30 sec, and 68^0C for 30 sec, and a final extension at 72^0C for 15 min. All PCR

assemblies were performed in the ancient DNA laboratory facility and amplifications were carried out in a separate and physically isolated molecular biology lab. Each PCR contained an extraction control and a no template PCR control. PCR products were visualized on 2% agarose gels stained with ethidium bromide. PCR products were TA-cloned and sequenced as described previously[16].

3.3 Results

Mitochondrial DNA (mtDNA) sequencing showed substitutions that characterize the mtDNA ofAnzick-1 individual as a member of sub-haplogroup D4h3a[17], which led us to proceed with full genome sequencing. More detailed discussion of the mtDNA results can be found in SI section 12.

3.4 Mitochondrial DNA Typing and Sequencing of Elk samples

The four DNA extracts from the presumed elk bones were analysed for the presence of cervid mitochondrial DNA. All pre-PCR laboratory work was conducted in dedicated ancient DNA facilities at Royal Holloway following standard protocols. Amplifications were conducted in 25μL reactions that comprised 1x PCR buffer, 1mM of additional MgCl2, 1mg/mL BSA, 200μM of each dNTP, 0.4μM of each primer, 1U HotStar Taq (Qiagen), 1μL of DNA extract, and purified water. Amplification conditions consisted of five minutes at 95°C, 50 cycles of one minute at 94°C, one minute at 47-56°C, and one minute at 72°C, followed by ten minutes at 72°C, and ten minutes at 12°C. Sequencing and data analysis protocols followed those outlined in Brace *et al*.[18] Initial testing for a 112bp fragment of the cytochrome b locus, amplified with a novel primer set (forward: 5-GAATCCCATCAGATGCAGACAAAA and reverse: 5-GAGTGGGTTTGCTGGGGTGTAG), demonstrated the presence of *Cervus elaphus* DNA in this specimen. To identify the mitochondrial haplotype to which this individual belonged, another 412bp section of cytochrome b and 430bps of the control region (CR) were amplified, using the multiple, overlapping primer sets of Meiri *et al.*[19,20] and two novel CR sets (5-TACTAATTACACAGCAAAAC and 5-TACATTAATTTATGTACTATTATACA; 5-AACAGTACATGAGTTAGCG and 5-ACATTATGTACTATGTACTTCAGA). Observed sequence mismatches were all C→T and G→A transitions, which are typical artefacts of *post mortem* miscoding lesions. Alignment to a large dataset of Siberian and North American *C. elaphus* sequences[20] revealed that the osseous tool sample belongs to a haplotype (MM004) that was distributed in East Beringia

(Yukon, Alaska) from 14,426 to 12,698 cal. years BP. Additionally, this haplotype is known from East Siberia (531 cal. years BP) and Alberta (undated)[20].

4. Extraction, library build and shotgun sequencing

Bone powder was obtained from 2 small cranial bone fragments and a small piece of rib from the Anzick-1 infant skeleton; between 50 and 100mg of powder was collected. DNA extraction was done as previously published[21,22]. Briefly, bone powder was dissolved in 5ml digestion buffer (0.47M EDTA, with 0.5% N-laurylsarcosyl and proteinase K) for 24-48hrs at 37°C. The digest was added to a buffered binding buffer (5M GuSCN, 0.05M Tris-HCl pH=8, 0.05M NaCl, 0.02M EDTA, 1% TritonX-100) with 100ul dissolved fine-grain silica[21] and final pH of 4.0-5.0 after adjustment with 32% HCl. After incubating 3hrs at room temperature, silica was pelleted and washed with ice-cold 80% ethanol twice, and eluted in 100µl of EB (Qiagen, Germany).

Library preparation was done using a commercial kit from New England Biolabs (E6070, Ipswich, MA) following manufacturer's guidelines, with the following modifications: all reactions where scaled to half-volume, no DNA fragmentation was performed, DNA purifications were done using MinElute columns and PN binding buffer (Qiagen, Germany). Adaptors were made as described in Meyer and Kircher[23], and used in a final concentration of 2.5µM each. A total of 19 libraries were produced from the Anzick-1 sample.

Library amplification was done in two steps, initial round of amplification utilizes indexed primers of the form, inPE1.0: 5'- AATGATACGGCGACCACCGAGATCTACACTCTTTC-CCTACACGACGCTCTTCCGATCT, indexPE2: 5'- CAAGCAGAAGACGGCATACGA-GATNNNNNNGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT, where the index is represented by 6N nucleotides. Second round amplification utilizes primers P5: 5'-AATGATACGGCGACCACCGA and P7: 5'-CAAGCAGAAGACGGCATACGA, all primer sequences are from Illumina (San Diego). For 4 libraries built early in the process (C7, C8, C9, C10, see Extended Data Table 3) amplification was done using AmpliTaq Gold, in the following setup for the initial PCR: 12.5µl DNA library, 1X AmpliTaq Gold buffer (Applied Biosystems, Foster City, CA), 2mM MgCl2, 3% DMSO, 0.4mM dNTP each (Invitrogen, Carlsbad, CA), 0.2µM primer inPE1.0 and indexPE2, 1 unit of AmpliTaq Gold (Applied Biosystems, Foster City, CA) and H2O to 50µl. This was amplified with the following cycling conditions, 10' at 94°C, 12 cycles of (30" at 94°C, 30" at 60°C and 30" at

72°C), with a final extension at 72°C for 10'. The second PCR was setup as a reamplification, with the same cycling conditions and reaction mix, with only the primers substituted for P5 and P7, the template was 5µl of the initial PCR. The remaining libraries were amplified using KAPA HiFi Uracil+ (Kapa Biosystems, Woburn, MA), for initial amplification we used 25µl 2X master mix, 0.2µM each primer, 3% DMSO (Invitrogen, Carlsbad, CA), 0.2mg/ml BSA (New England Biolabs, Ipswich, MA) for 12.5µl of unamplified library in a final volume of 50 µl. This was amplified with the following cycling conditions, 4' at 98°C, 8 cycles of (10" at 98°C, 30" at 62°C and 30" at 72°C), with a final extension at 72°C for 10'. The second PCR was setup as a reamplification, with the same cycling conditions and reaction mix, with only the primers substituted for P5 and P7, the template was 5µl of the initial PCR. The number of cycles for the secondary PCRs were adjusted to give a final concentration between 5 and 20nM, in the size range from 130bp and above, as quantified on a BioAnalyzer High Sensitivity assay. Final PCR products were pooled and submitted for 100 bp single end sequencing at the Danish National High-Throughput Sequencing Facility.

5. Additional modern reference genomes

DNA from the Mayan (HGDP00877) and Karitiana (BI16), were collected from CEHG collection and the Cavalli-Sforza collection, respectively. The Mayan sample was sheared using Covaris, and build into DNA libraries with TruSeq-kit (Illumina, San Diego) following manufacturers guidelines. The Karitiana sample was processed using the Nextera library kit (Illumina, San Diego) following suggested guidelines. Genomic amplifications were carried as specified by vendor for up to 7 cycles. Both samples were submitted to sequencing at the Stanford Genome Center and 3 full lanes per sample were sequenced on the HiSeq2000 platform (Illumina, San Diego).

6. Mapping and genotyping

6.1 Read data processing

The Illumina data were basecalled using Illumina software CASAVA 1.8.2 and sequences were de-multiplexed with a requirement of full match of the 6 nucleotide index that was used for library preparation. Samples prepared using Nextera (HGDP00877 and BI16) were hard clipped 13 nt of the 5'. Adapter sequences and leading/trailing stretches of Ns were trimmed

from the reads and additionally bases with quality 2 or less were removed using AdapterRemoval-1.1[24]. Trimmed reads were mapped to the human reference genome build 37 and hg18 using bwa-0.6.2[25] and filtered for mapping quality 30 and sorted using Picard (http://picard.sourceforge.net) and samtools[26]. Data from ancient samples were mapped using the seed disabled to allow for better sensitivity[27]. Data were merged to library level and duplicates removed using Picard MarkDuplicates (http://picard.sourceforge.net) and hereafter merged to sample level. Sample level BAMs were re-aligned using GATK-2.2-3 and hereafter had the md-tag updated and extended BAQs calculated using samtools calmd[26]. Read depth and coverage were determined using pysam (http://code.google.com/p/pysam/) and BEDtools[28]. The sequence reads and alignments were available at the Sequence Read Archive (SRA) under the accession (SRX381032).

6.2 Genotyping

All samples were genotyped using samtools-0.1.18 mpileup and bcftools[26]. Each sample was genotyped individually and filtered/masked to achieve a high confidence call set similar to Ragahavan et al.[29] In brief, the calls were filtered for: strand and distance bias; phred-scaled genotype posterior probability > 20; read depths between 10-100 for autosomes, 5-50 for Y and X (except 10-100 for X in females) and 10- for MT; not within an annotated repeat; allelic balance greater than 0.2 for heterozygote sites; not within 5 nts of another variant; and heterozygote calls on X, Y and MT masked for males and MT for females. Calls were combined using GATK CombineVariants 2-5.2[30] and filtered for sites that violated a one-tailed test for Hardy-Weinberg Equilibrium at a p-value < $1e^{-4}$ [31]. The variants are available for download from http://www.cbs.dtu.dk/suppl/clovis/.

6.3 Phasing and ancestry painting

Variants sites from the 20 individuals were phased using Beagle4 (Browning and Browning, 2013) using the 1000 Genomes Project phase 1 reference panel. To identify regions of the genome with putative affinity to European, Native American, Asian or African populations we used a discriminative approach as implemented in the program RFmix[32]. The method allows to model the ancestry of each phased individual chromosome of the Anzick-1 individual using known ancestries from a reference panel comprised of phased genotyped individuals of European (n = 43), Native American (n = 43), Asian (n =43) and African origin (n = 43). As reference panels we used the HapMap project[33] with the Affy 6.0 individuals identified as European, Native American, African and Asian. The phased Anzick-1 individual

was intersected with the reference panel rendering a total of 394227 SNPs (chr1: 31969, chr2: 34477, chr3: 27433, chr4: 24645, chr5: 25854, chr6: 25834, chr7: 21544, chr8: 21647, chr9: 18218, chr10: 22594, chr11: 19488, chr12: 18640, chr13: 15483, chr14: 13253, chr15: 12432, chr16: 12016, chr17: 9887, chr18: 12635, chr19: 5015, chr20: 10174, chr21: 5833, chr22: 5156). We recognize that performing local ancestry analysis on diachronic samples poses additional problems of interpretation when compared to analysis performed on just modern genomes. Genetic drift will contribute to the differences observed in modern samples from different geographic locations as well as the differences between modern and ancient samples from the same location. We expect the importance of this effect to increase as the reference populations employed are more closely related; thereof our decision to select reference populations of modern genomes presenting Asian, European, African and Native American ancestries. These ancestries will present enough differentiation to facilitate our understanding of the pattern of genetic diversity in the Anzick genome when compared to current Native American genomes. Global proportions of shared ancestries were estimated using Admixture[34] on the dataset containing the intersected set of SNPs for the reference panels and Anzick-1 individual. We used cross-validation errors to decide the number of clusters (putative ancestries) that better explain the data, and decided on 4 ancestral clusters as the best set-estimate. The estimation of standard errors in the assignment of global ancestries was performed via bootstrap as implemented in the Admixture software. The Anzick-1 individual shares most ancestry with Native Americans with small patches shared with Asians and less with Europeans.

The same analyses were performed on the modern Native American genomes generated for this study. While the two Karitiana genomes showed very high Native American ancestry, the Mayan individual showed a proportion of European ancestry of 12% (s.e. 1%). Based on this observation, we chose to mask the Mayan individual for European ancestry by removing any sites for which the individual had at least one of its alleles identified as European. There is roughly a 7% difference between the proportions of Native global ancestry assigned to the Anzick-1 genome and the modern genomes. It is highly likely that the differences observed in the ancestry painting of current Native American genomes and the Anzick-1 individual are due to temporal drift that has caused the ancient Native American sample to differentiate from the modern ones. Although we cannot specifically talk about the contribution of Native American ancestry from the Anzick-1 individual on the modern genomes or vice versa, we can certainly interpret the shared ancestries between modern Kairitianas and Mayans and the

ancient individual as belonging to the same original population that gave rise to both populations, potentially the ancestral Clovis population that later expanded into Central and South America.

7. DNA preservation in the Anzick-1 fossil

To assess the level of DNA preservation of the Anzick-1 sample, we examined the DNA fragment length distribution of all reads mapping to the human reference genome (Extended Data Figure 1a).

In an ancient sample random chain scission will result in a negative exponential correlation between the number of DNA molecules and their length. This reflects *post mortem* DNA fragmentation (facilitated mainly by depurination) leaving few long DNA fragments and many short ones. In the Anzick-1 sample there was a surprisingly large proportion of DNA fragments >94 bp (Extended Data Figure 1a), which is our maximum sequencing length. This is indicative of exceptional DNA preservation despite the geologic age of the sample.

Following Allentoft et al.[35], we examined only the declining part of the fragment length distribution (Extended Data Figure 1a), knowing that the decreasing number of shortest fragments is a result of these being lost in the DNA extraction. As demonstrated in Deagle et al.[36], the decay constant (λ) in this exponential relationship represents the damage fraction (i.e. the fraction of broken bonds in the DNA backbone). Solving the equation for nuclear DNA and mtDNA, respectively, we retrieved DNA damage fractions (λ) of 1.8 % and 1.7 %. The observed difference corresponds with previous findings that mtDNA are being cleaved at a slower rate than nuclear DNA[35], perhaps owing to its circular structure and/or protective properties of the double membrane of mitochondria or both.

It has been shown that long-term *post mortem* DNA fragmentation can be described as a rate process, and that the damage fraction (λ, per site) can be converted to a decay rate (k, per site per year), when the age of the sample is known[35]. The 95.4% probability for the calibrated age of the Anzick-1 sample spans 12,722-12,590 years BP. With "present" being 1950 AD, we use 12,785 years (12,722 + 63) as the likely time period between the death of the individual and the DNA extraction in 2013. The corresponding decay rates (k) were 1.43E-6 and 1.35E-6 for nuclear DNA and mtDNA, respectively. We calculated the molecular half-life ($t_{1/2} = \ln 2/k$) in the Anzick-1 fossil to 4852 and 5137 years for a 100 bp nuclear DNA and

mtDNA respectively. After this time, half the fragments of this length will be gone due to one or more strand breaks.

7.1 Predictions

It is evident from the results above, and the relatively high endogenous DNA content (up to 28.2%) that the DNA preservation in the Anzick-1 fossil is extremely good despite its age. This is very likely owing to a low ambient burial temperature at the Anzick site in Montana (see below). A general model has been proposed for the temperature-dependency of the rate of mtDNA decay (k) in bone[35]:

$$\ln k = 41.2 - 15267.6 * 1/T,$$

where T is the absolute temperature.

We tested if this model could describe the observed DNA degradation of the Anzick-1 sample. With an effective burial temperature estimated to 4.4°C (see below), the predicted mtDNA damage fraction (λ) was 1.3% and the rate (k) was predicted to 1.02E-6 strand breaks, per site per year. With a predicted value remarkably close to the k of 1.35E-6 estimated for the mtDNA, the model confirms that the DNA in the Anzick-1 sample should be extremely well preserved, providing support for the authenticity of the DNA identified in this sample.

In line with previous findings[35], we note that the DNA decay rate from this 'shot-gun' sequencing data is slightly faster than the predicted value. This is likely because the decay rate and proposed model in Allentoft et al.[35] was based on quantitative PCR data, and relied on DNA copy number differences between ancient samples of different ages. This implies that the model is restricted to predict the rate of long-term DNA degradation, which is a pure chemical reaction (depurination). In contrast, decay rates calculated from fragment length distributions (i.e. Anzick-1) will also incorporate the initial *post mortem* (and likely much faster) DNA degradation phase, facilitated by enzymatic activity. Hence, when averaged over time, the observed overall rate will be faster than predicted from the model. As more empirical ancient genomic data become available, we will be able to assess this in greater detail and refine the model accordingly. Regardless, this dataset adds to a growing body of evidence, which suggests that it is possible to produce reasonable predictions of the DNA preservation in a sample, when the age and burial temperature are known.

7.2 Estimating the effective burial temperature (T_{eff})

We estimated the effective burial temperature (T_{eff}) for the Anzick-1 fossil to 4.4°C (www.thermal-age.eu; analysis S693). This estimation assumed that despite a vertical depth of 2.5 m in the talus slope, it would have lain within 1 m horizontally of the slope's surface in sediment having a net thermal diffusivity of 0.029 m^2 day^{-1} (silt-loam, 10% water) for 12,740 years (time until excavated in 1968) and then in highly variable conditions (15°C±7°C) for 35 years until present day. Seasonal fluctuations in monthly temperature at the site (45.987 °N - 110.653 °E) were estimated from the WorldClim[37]. The altitude difference between the 1 x 1 km squares of WorldClim and the site was corrected by comparing the altitude of the WorldClim grid with the altitude from site reports (5,000', 1524 m) and corrected using a standard environmental lapse rate of 6.49°C/1,000m. The extent to which temperature decreased beyond the Holocene was estimated from the difference in the 1° × 1° PIMP2 grid for the region at three time intervals Modern (pre-industrial), Holocene (6ka) and LGM[38]. These data were correlated against the equivalent time internals from the Bintanja et al.[39] curve, and the correlation was used to transform the latter temperature series to reflect the temperature change in this region.

We do not have detailed data on the net thermal diffusivity of the soil, original burial depth, burial depth over time, or storage history since 1968 AD. We also note that the model is very sensitive to temperature, and this is illustrated by the fact that if the burial depth is increased from 1 to 2.5 m, then the estimated decay rate falls by 40% (k = 0.628E-6 per site, per year, S696).

8. DNA damage patterns

Ancient DNA has been shown to accumulate DNA damage mainly due to deamination of cytosine, giving rise to a characteristic pattern, showing an increased C to T rate at the 5' end of the read, and similarly G to A on the 3' end of the read[40]. To test for this pattern in the Anzick-1 individual, we randomly selected a subset 0.5% uniquely mapped reads and performed mapDamage2 analyses[41]. The Anzick-1 individual shows a clear nucleotide misincorporation signature of DNA damage (Extended Data Figure 1b). Additionally, fragmentation patterns revealed an increase in purines at the genomic coordinate located upstream of sequencing starts, in agreement with depurination being one of the key-drivers of post-mortem DNA fragmentation[40,42]. Nucleotide misincorporation and DNA fragmentation

patterns were found asymmetric at read starts and read ends due to the fact that a significant fraction of the DNA templates was not sequenced over their full length. Overall, we observed typical signatures of DNA damage, which suggests that the sequencing data analysed originates from ancient DNA templates and not modern DNA contaminants.

9. mtDNA consensus and contamination estimate

As described in (SI6), the raw reads were first trimmed using AdapterRemoval-1.1 (Lindgreen 2012) for adapter sequence and leading/trailing Ns to a minimum length of 25 nt (--minlength 25 --trimns). We then used two different mapping strategies; one to call the consensus sequence and one to determine the contamination fraction. In both cases the mapping was done using bwa-0.6.2[25] with seed disabled to allow for better sensitivity[27]. Alignments were filtered for reads with a mapping quality of at least 30, sorted and merged to libraries using Picard (http://picard.sourceforge.net). Library BAMs had duplicates removed using Picard MarkDuplicates, were merged to sample level and realigned using GATK[30]. Last, the md-tags were recalculated using samtools. We visualized the results of the mapping using tablet v.1_13_07_31[43].

9.1 mtDNA consensus sequence

To determine the consensus sequence, we proceeded with a two-step mapping iteration to avoid biasing the result by the reference sequenced used. We first mapped the reads to an mtDNA genome to determine the consensus (as opposed to the nuclear genome plus the mtDNA genome) since it has been shown that the nuclear inserts are in low enough frequency to call a consensus sequence for the mtDNA[44].

For the first iteration, we mapped the reads to the revised Cambridge reference sequence (rCRS)[45] as described above. The average depth after mapping to the rCRS is 185.83 while 100% of the molecule is covered by at least one read. We then used samtools[26] to call the consensus by discarding bases with a base quality lower than 20. All insertions and deletions were then checked manually. In the next step, we used blast[46] (default parameters) to find the closest modern sequence publicly available, which was found to be the sequence of a native American individual (Genbank: FJ168755.1, haplogroup D4h3a). We then mapped the original reads against this new sequence and called a new consensus. We mapped the reads against that consensus and found that the iteration had converged.

Several approaches were applied to evaluate the quality of the final consensus. First, we checked that all positions had a major allele frequency greater than 80% and additionally found that only 12 positions had major allele frequency lower than 90% and they all involved C>T and G>A substitutions characteristics of ancient DNA damage[40]. Second, all positions were covered by at least one read that did not contain any mismatches relative to the consensus, and the number of such reads was greater or equivalent to that of the read with 1 mismatch (relative to the consensus) at all positions, which is expected under the scenario that the right consensus is called.

9.2 Contamination estimates

For contamination estimates, once the consensus sequence was determined, we mapped the reads against the entire build37.1 augmented with the consensus, to reduce the problem of nuclear inserts that would look like contaminants.

To estimate contamination, we used a method detailed in the Supplemental Information section 5 of[47] that generates a moment-based estimate of the error rate and a Bayesian-based estimate of the posterior probability of the contamination fraction.

The method is based on a Monte Carlo Markov Chain (MCMC) method. We ran three independent chains for 50,000 iterations discarding the first 218. Visual inspections of the shrink factor over every iteration of the chains as well as a potential scale reduction factor of 1 suggest that the MCMC has converged.

We found the 95% credible interval of the posterior probability of contamination rate to be (0.15%,0.38%) while the mode of the posterior probability is 0.25%.

10. Error estimation

10.1 Brief description of the method

We estimated the error rates with a method similar to the method used by Reich *et al.*[48] that makes use of a high quality genome. The estimation is based on the idea that any given human sample should have the same expected number of derived alleles compared to the chimp sequence. We estimate the expected number of derived alleles from the high quality genome and assume that any excess of derived alleles observed in the sample individual is due to errors. If the high quality genome has no errors then the error rate estimate of the sample is equal to the true error rate. However, if the high quality genome does have errors,

the estimated error rate can roughly be understood as the excess error rate relative to the error rate of the high quality genome.

We estimated overall error rates using a method of moment estimator, while type specific error rates were estimated based on a maximum likelihood approach. The model and the estimation methods are described in details in Orlando *et al.*[49], SI text 4.4. Note that unlike Orlando *et al.*[49] we used all reads instead of sampling a single read per site.

10.2 Data

We made estimates both for the entire Anzick-1 sample and for each individual library separately.

For the chimpanzee we used the multiway alignment that includes both chimpanzee and human (pantro2 from the hg19 multiz46). For the high quality genome we used sequencing data for the individual from the 1000 genome with ID NA06985 and excluded all reads with a mapping quality score less than 30 and all bases with a base quality score less than 20. The same quality filters were used for all the data we estimated error rates for.

10.3 Results

The estimated error rates for the sample (Anzick-1) and each library can be seen in Extended Data Figure 1c.

11. X chromosome based contamination analysis

11.1. Brief description of the methods

Based on the sequencing data that mapped to the X-chromosome we inferred the contamination rate of the male ancient sample using the methods described in Rasmussen *et al.*[50]. These methods exploit that since the X chromosome is haploid for male individuals any discordance in observed bases in a single site is either due to sequencing errors or contamination, which means that discordance rates in the X chromosome contain information about contamination.

The methods are based on a fixed set of SNPs known to be polymorphic in Europeans. We assume that regardless of population, the probability of a site being polymorphic is higher for the set of known polymorphic sites compared to the adjacent sites. We also assume that the error rate of the set of known polymorphic sites is the same as the error rates of the adjacent

sites. Under these assumptions the base discordance rate for the known polymorphic sites should be the same as for the adjacent sites if there is no contamination. In contrast, contamination from any human source will lead to a higher impact on the discordance rate for the known polymorphic sites than the adjacent sites. Hence by comparing discordance rate in known polymorphic sites to the discordance rates in their adjacent sites the amount of potential contamination can be tested and estimated. We use two approaches for this: "Test 1" in which all reads are used and "Test 2" in which only a single sampled read is used. Test 2 is less powerful but does not assume that the errors are independent between reads and sites. For details and validation of the methods, see Rasmussen *et al.*[50]

11.2. Data

The set of known polymorphic sites was identified using 60 unrelated CEPH individuals from the HapMap phase II release 27 data[51]. This set was pruned such that no polymorphic sites were less than 10 bases apart. Based on these 60 individuals, we also estimated the allele frequencies in the European population.

For the Anzick-1 genome the following filtering was performed:

- The X chromosome was trimmed to remove the regions that are homologous with the Y chromosome (first and last 5Mb).
- The sites were then filtered based on mappability (100mer), so that no region will map to another region of the genome with an identity above 98%
- Reads with a mapping quality score of less than 30 and bases with a base quality score less than 20 were removed.
- Sites with a read depth of less than 4 or above 40 were removed.

11.3. Results for Anzick-1

Both Test1 and Test2 showed evidence for low presence of contamination; *p*-values for both tests, obtained using Fisher's exact test, are below 10^{-10}. The amount of contamination was estimated to 0.86% (s.e. 0.059%) and 1.2% (s.e. 0.18 %) for Test1 and Test2, respectively.

12. mtDNA analyses

Methods of Anzick-1 mtDNA analysis

Haplogroup affiliation of the mitochondrial genome of Anzick-1 specimen was determined based on the nucleotide position differences from the reference sequence (NCBI Reference Sequence: NC_012920.1) by the use of the software programs FASTmtDNA and mtDNAble, provided by mtDNA Community (www.mtdnacommunity.org). The schematic phylogenetic tree of mtDNA haplogroup D4h (Extended Data Figure 2a), showing the location of the Anzick-1 mtDNA haplotype within present-day variation was built according to present nomenclature of mtDNA Tree Build 15, www.phylotree.org[17], using the summarized data from[52-55].

12.1 Results of Anzick-1 mtDNA analysis (Extended Data Figure 2a)

MtDNA of the Anzick-1 genome belongs to haplogroup D4h3a, one of the rare mtDNA haplogroups specific to Native Americans and is distributed today along the Pacific coast in North and South America. Its sister-branch D4h3b consists of only one D4h3 mtDNA lineage found in Eastern China[56]. It has been proposed that D4h3a was introduced into the Americas during the early colonization process, and that it spread rapidly southward following the Pacific coastal route, with subsequent increase in frequency in the Americas[53].

The D4h3a branch [with an age estimate of 13,000 +/- 2600 years[57]] encompasses all Native American D4h3 mtDNAs (including mtDNA of Anzick-1) and has been mostly found in South America (Chile, Peru, Ecuador, Bolivia, Brazil), and to a much lesser extent also from Mexico and California (see the spatial distribution of D4h3a in Fig. 3 in[53]). The highest frequency of D4h3a has been found from Cayapa Amerinds from Ecuador (22%)[58]. D4h3a has several sub-branches (D4h3a1-7)[17], the genome of the Anzick-1 specimen does not have defining mutations of these sub-branches and is located in the root of D4h3a (Extended Data Figure 2a), thus being ancestral to all modern D4h3a haplotypes. Although D4h3a is rare in modern North American populations, the evidence of its early presence among Native Americans in North America is the Southeast Alaska 10,300-year-old skeletal remains that were found to have the corresponding to D4h3a control region motif in its mitochondrial genome[59]. Similarly, the full sequence of another ancient sample (ca 6000 YBP) from the Northwest Coast of North America belongs into hg D4h3a[55] (Extended Data Figure 2a). The control region motif of D4h3a has been found also from the much younger skeletal populations (dated to 100-400 YBP) from Tierra del Fuego in South America[60] and at the Klunk Mound cemetery site in West-Central Illinois 1800 years BP[61].

13. Y-chromosome analyses

We merged Y-chromosome reads with 29 previously analysed sequences from: 15 Human Genome Diversity Panel[62] samples, 2 Gabonese individuals[63], an ancient Saqqaq Palaeo-Eskimo[64] and all 11 haplogroup Q (hg Q) samples from phase 1 of the 1000 Genomes Project[65]. We called genotypes jointly among the 30 individuals across approximately 10 Mb of the Y chromosome, as described in Poznik *et al.*[63]. We then used MEGA[66] to construct a bootstrap consensus maximum likelihood tree, and we assigned each variant site to a specific branch. In Extended Data Figure 2b, we present the P-M45 subtree, represented by 16 carriers of the M45 SNP. In previous analyses[63], three hg Q-L54*(xM3) individuals shared a branch representing 21 total derived alleles (*i.e.*, including transitions; branch #25), and Anzick-1 carried 18 of these, including the well-known L54 SNP (3 sites had missing data). In addition, we observed 13 reads of support for the ancestral G allele at marker M3. Thus, Anzick-1 is a member of Y-chromosome haplogroup Q-L54*(xM3). This observation is in line with expectation, as haplogroup Q and subgroup Q-L54 originated in Asia, and two Q-L54 descendants predominate in the Americas: Q-M3, which has been observed exclusively in Native-Americans[67] and in Northeast Siberia[68], and Q-L54*(xM3).

With perfect preservation, one would expect the terminal branch leading to Anzick-1 to be quite a bit shorter than the others due to the fact that mutations have had ~12,600 fewer years to accumulate. However, when all SNPs were considered, this reduction was more than counter-balanced by post-mortem transition mutations, which had significantly lengthened the branch. However, when we restrict our attention to transversion SNPs, as in Extended Data Figure 2b, and to the four hg Q samples for whom the false negative rate is minimized due to haploid coverage of at least 5× (HGDP00877, HGDP00856, HG01124, and Anzick-1), we can compute a rough estimate for the time to the most recent common ancestor (TMRCA) of the Q-M3 and Q-L54*(xM3) haplogroups. The modern samples have accumulated an average of 48.7 transversions since their TMRCA, and we observed 12 in Anzick-1: 5 unique and 7 shared. We infer an average of approximately 36.7 (48.7 – 12) transversions to have accumulated in the past 12.6 ky and therefore estimate the divergence time of Q-M3 and Q-L54*(xM3) to be approximately 16.8 ky (12.6 ky × 48.7 / 36.7). Since this intuitive approach does not combine information from the modern lineages in the most informative way, we implemented a Poisson process model for mutations on the tree and used the `constrOptim()` function in R to compute a maximum likelihood TMRCA estimate of 16.9 ky. We then repeated this for 100,000 bootstrap simulations to yield a 95% confidence

interval of 13.0–19.7 ky. Anzick-1 had a missingness rate of 0.10. Accounting for this would add approximately 0.6 singleton transversions and have little impact on the TMRCA calculations.

Interestingly, the Palaeo-Eskimo Saqqaq lineage constitutes an outgroup to the Q-L54 ensemble composed of all other Native-American hg Qs within the sample. The individual carried the NWT01 SNP, which characterizes approximately half of modern Inuvialuit speakers from the Canadian Northwest Territories[67], indicating a continuity of this lineage to the present day. Because the Saqqaq sequence had a relatively high missing rate of 0.24 and is divergent with respect to the other hgQ lineages in the sample, its singleton branch should more properly be considered to be of length 71 (54 / 0.76) transversions, and it is likely that approximately four transversions assigned to the Q-L54 branch (#26) are in fact possessed by Q-NWT01 lineages. With more Q-NWT01 sequences in the sample these SNPs could be identified and re-assigned to branch 28.14. Clustering analysis of the Anzick-1 individual and 1803 genotyped worldwide individuals

14. Clustering analysis of the Anzick individual and 1803 genotyped worldwide individuals

14.1 Methods

We next used a STRUCTURE-like[69] but maximum likelihood based approach assembled into ADMIXTURE[34] to reveal the relationships of the Anzick-1 genome to modern human genetic diversity in Eurasia and the Americas. We compiled a reference SNP array data set starting from 2,081 individuals curated and assembled from a number of studies[51,64,70,71] by Reich *et al.*[72] In this data set genomic regions in Native American and Siberian populations which were found to be of recent African or European ancestry were identified and excluded[72]. This data set was merged with additional data from 401 individuals genotyped on Illumina arrays by several recent studies[29,73-77], obtaining a final total of 331,710 SNPs. We obtained haploid genotype calls from the Anzick-1 individual for genomic positions present in this modern-day reference data set by first excluding all sequence reads with a phred-scaled mapping quality less than 30 and all bases with a phred-scaled mapping quality less than 30. We next required support from at least 3 sequence-reads for an allele to be called. If more than one such allele was present, we randomly sampled one for the haploid call. If this allele was not one of the two alleles present in the SNP array data, we excluded the haploid call. While transition SNPs

are sensitive to post-mortem damage in ancient DNA, we include all SNPs in the analyses since the reference data set contains very few transversion polymorphisms. We excluded all positions at which the Anzick-1 individual had a third allele not observed in the modern-day individuals. To allow comparisons between the haploid data from the Anzick-1 individual and modern-day individuals, we randomly sampled one haploid allele from each modern-day individual, effectively haploidizing all individuals in the analysis.

We ran ADMIXTURE assuming different number of clusters in 100 replicates. We monitored convergence of individual Admixture runs at each K by looking at the maximum difference in Log Likelihood (LL) scores in the fractions of runs with the highest LL scores at each K. We report the runs for K=3 to K=5 and K=9 to K=11 (Extended Data Figure 3). We note that the K number of clusters that is found is relevant only considering the populations in the specific sample set used.

14.2 Results and Discussion

At K = 11 the Anzick -1 genome consists of five major genetic components which are the same five that constitute the extant Native American genetic variation (Extended Data Figure 3). It is important to note that the five components do not imply five ancestral populations for Native Americans. Instead this analysis reveals that the genetic structure in the Native American populations included in this setting considering also the background of the included Eurasian populations, is best described in terms of five components. Importantly, from this analysis we are not able to tell neither the nature (population splits; isolation by distance; admixture) nor time of the demographic events leading to the patterns of observed genetic structure. The key observations are nevertheless relevant. The Anzick-1 genome has all of the components present in contemporary Native Americans, the most predominant genetic components in the Anzick-1 genome are the same that are dominant in most Central and Southern Native Americans today.

15. Analyses using outgroup f_3- and *D*-statistics of the Anzick-1 individual and modern-day worldwide populations

15.1 Data preparation and processing

We obtained haploid genotype calls from the Anzick-1 individual for genomic positions present in modern-day reference data sets typed on large-scale SNP arrays (see below). We

excluded all sequence reads with a phred-scaled mapping quality less than 30 and all bases with a phred-scaled mapping quality less than 30. To alleviate the effect of sequence errors and post-mortem nucleotide misincorporations in the data from the Anzick-1 individual, we required support from at least 3 sequence reads for an allele to be considered. If more than one such allele was present, we randomly sampled one for the haploid call. If this allele was not one of the two alleles present in the SNP array data, we excluded the haploid call.

Except when otherwise noted, a published SNP dataset from 2,081 individuals assembled from a number of studies was used [33,51,64,70-72]. The data set we used was curated by Reich et al.[72] who also identified and excluded genomic regions in Native American and Siberian populations, which were found to be of recent African or European ancestry. The data set was merged with additional data from 81 Finnish individuals[73], resulting in a final total of 335,461 SNPs with strand orientation according to the hg18/NCBI 36 human genome reference assembly. While transition SNPs are sensitive to post-mortem damage in ancient DNA, these data contained very few transversion polymorphisms, therefore we opted to include all SNPs in the analyses (unless otherwise noted).

For the outgroup f_3 analyses we added additional individuals from Eurasia to this data set to increase the resolution for the mapping Old World affinities to the Anzick-1 individuals. We thus merged the data above with data from 320 individuals from several recent studies[29,74-77], and obtained a total of 331,710 SNPs with strand orientation according to hg19/NCBI build 37.1.

For some analyses we also used 269,441 SNPs in hg18 strand orientation that were cleanly ascertained in either a San or Yoruba individual. Since the San and Yoruba are approximately outgroups to non-African populations, these data are unbiased for all comparisons between non-Africans[78]. For these tests, the 12 individuals that were used for ascertainment by Patterson et al.[78] were always excluded.

15.2 Outgroup f_3 statistics reveal that the Anzick-1 individual is most closely related to Native Americans

To obtain a statistic that is informative of the genetic relatedness between a particular sample and each modern population in a reference set, we computed an 'outgroup f_3-statistic'[29], where the deviation from 0 will be a function of the shared genetic history of two populations A and B in their unrooted history with the outgroup O. We used the estimator suggested by Patterson et al.[78]

$$f_3(A,B;O) = \frac{\sum_{i=1}^{n}[(p_{iO}-p_{iA})(p_{iO}-p_{iB}) - \left(\left(\frac{h_{iO}(k_{iO}-h_{iO})}{k_{iO}(k_{iO}-1)}\right)/k_{iO}\right)]}{\sum_{i=1}^{n}[2p_{iO}(1-p_{iO})]}$$

where h_{iO} is the count of the reference allele (arbitrarily chosen between the two alleles present at locus *i*) and k_{iO} is the number of gene copies in population *O* (the outgroup) at locus *i*, with corresponding notations for populations *A* and *B*.

We computed this statistic with the Anzick-1 individual as population *A*, one of 143 modern-day populations as *B* and West African Yoruba as *O*, and found that the Anzick-1 individual is clearly most closely related to modern-day Native American populations. However, this affinity is weaker in 7 of the 52 Native American populations in the data: Aleutians, East Greenlanders, West Greenlanders, Chipewyan, Algonquin, Cree, and Ojibwa (Figure 2a). In Eurasia, genetic affinity with the Anzick-1 individual decreases with distance from the Bering Strait (Figure 2a). To confirm that the Anzick-1 individual is most closely related to Native Americans, we repeated the analysis with cleanly ascertained SNPs in San and Yoruba typed in 52 modern worldwide populations, which demonstrates that the results here are not due to ascertainment bias (data not shown).

15.3 The Anzick-1 individual shares more recent history with Central and South Americans than Northern Amerinds and a Yaqui individual from Mexico

A lower degree of genetic affinity between the Anzick-1 individual and Aleutians, East Greenlanders, West Greenlanders, and Chipewyan is consistent with previous suggestions of these populations carrying >10% ancestry from secondary waves of gene flow from Asia into the Americas[72]. Our data suggest that these waves may postdate the time when the Anzick-1 individual lived. However, a previous large-scale study testing for the presence of multiple waves of migration into the Americas found the ancestry of the three Northern Amerind-speaking populations Algonquin, Cree, and Ojibwa to be entirely from the same migration that gave rise to Central and Southern American populations[72]. Since our outgroup f_3-statistics suggest that the Anzick-1 individual is less closely related to the Northern Amerind-speaking populations in our data, we explored the relationship between Anzick-1 and these populations further by computing *D*-statistics[78],

$$D(A,B;X,Y) = \frac{\sum_{i=1}^{n}[(p_{iA}-p_{iB})(p_{iX}-p_{iY})]}{\sum_{i=1}^{n}[(p_{iA}+p_{iB}-2p_{iA}p_{iB})(p_{iX}+p_{iY}-2p_{iX}p_{iY})]}$$

where p_{iA} is the frequency of one allele (arbitrarily chosen from the two alleles present) in population *A* at marker *i* and the statistic is summed for all *n* markers. This test is a generalization of the specific case when one gene copy is sampled from each of the populations *A*, *B*, *X* and *Y*, which makes up the sequence based test described in SI17. We obtained standard errors using a block jack-knife procedure over 5 megabase blocks in the genome[79].

We first tested if the Anzick-1 individual is as related to the South American Karitiana as each of the 51 other Native American populations (*X*) in the data using the statistic *D*(Han, Anzick-1; Karitiana, *X*). We found that this hypothesis could be rejected not only for the Aleutians, Greenlanders, and Chipewyan but also for Algonquin, Cree, Ojibwa and a Yaqui individual (Figure 2b). To confirm that this is not dependent on using the Karitiana as comparison, we performed the analogous tests *D*(Han, Anzick-1; Central/South Americans, Algonquin/Cree/Ojibwa/Yaqui) for all 44 Central South Americans other than the Yaqui (Extended Data Figure 4). We found that the pattern was consistent for all other populations than the Karitiana, but that some tests involving the Yaqui, Chorotega, and Kaingang where non-significant (likely due to both low sample size and a significant fraction of the genomes of these individuals masked for European ancestry[72]).

Moreover, we computed outgroup f_3 statistics as above, but replaced the Anzick-1 individual with pooled data from Algonquin, Cree and Ojibwa. This analysis contrasts the shared genetic history between Anzick-1 and a modern group, and the Northern Amerind-speakers and the same modern group (Extended Data Figure 5). We found that Central and South American populations all share a closer genetic history with the Anzick-1 individual than they do with modern-day North American individuals (Native American groups are skewed towards Anzick-1 from the 1:1 baseline). This supports the results above of a shared genetic history between the Anzick-1 individual and Central and Southern American groups that is not shared with the more Northerly groups in the data set.

15.4 The data are consistent with a tree-like model where Anzick-1 is ancestral to Central/South Americans

To investigate different models of population history that could have given rise to the differences in relatedness to the Anzick-1 individual and different Native American groups, we pooled all 3 Northern Amerind populations (“NA”; Algonquin, Cree, Ojibwa) and also data from 44 Central and Southern Native American populations (“SA”). To confirm that

these data mirrored the results above we first tested the hypothesis that Anzick-1 is basal to both "Northern" and "Southern" populations using Han Chinese as an outgroup. We obtain *D*(Han, Anzick-1; SA, NA) = 0.046±0.0046 (Z = 10.1), reiterating the closer relationship between the Anzick-1 individual and Southern Native Americans (Figure 3a).

Since we can reject a tree model where the Anzick-1 individual is ancestral to both Northern Native Americans and Southern Native Americans, we suggest three main models that could explain this pattern:

(1) The NA groups carry ancestry from a previously undocumented stream of gene flow from the Old World.

(2) The ancestors of Northern NA groups became isolated from the population lineage leading to both the Anzick-1 individual Central/Southern American groups, with little or no gene flow between NA and SA groups following the death of the Anzick-1 individual (Figure 3c).

(3) The Anzick-1 individual is from the ancestral population lineage of both NA and SA groups, which diverged only after the death of the Anzick-1 individual. After their divergence, the NA groups received gene flow from a Native American population lineage basal also to the Anzick-1 individual and are today carrying mixed ancestry both from the "basal" and SA lineages (Figure 3b).

To distinguish between these three models, we first computed tests of the form *D*(Yoruba, *X*; NA, SA) using 19 Siberian populations. The most negative Z-score we observe (negative values indicating affinity to the NA groups) was *D* = -0.0016 ± 0.002 (Z= -0.76) for Tundra Nentsi. The only significant ($|Z| > 3$) evidence for gene flow in these tests was from the SA lineage into the Naukan (*D* = 0.0080 ± 0.0015, *Z* = 5.19), which is consistent with the findings of Reich et al.[72] There is no evidence for Siberian gene flow into the Northern Amerinds consistent with the findings of Reich et al.[72] using the same Native American data and a very similar approach.

Next, we test the tree-like topology posited by model 2, and expect that the gene flow posited by model 3 (modern-day NA groups carrying "basal"+SA ancestry) would result in a deviation from tree-ness. However we find that the tree-like topology in model 2 is consistent with the data [*D*(Han, NA; Anzick-1, SA) = 0.005±0.006, Z = 0.87]. Thus, the data support a simple tree-like model where modern-day Southern and Central Native Americans are more closely related to the Anzick-1 individual than they are to modern-day Northern Native Americans. However, it is possible that the data could be explained by a model where the

Anzick-1 individual was contemporaneous with the common ancestral population of the NA lineage and the SA lineage (Figure 3d), after which the Northern Native American population received gene flow from a more basal Native American lineage. Note that the basal lineage must have been more closely related to Native Americans than Asian and Siberian groups, as the pattern that we observe in Northern Native Americans is not detected using Siberian groups alone (see above and [72]). This model would thus still imply an early divergence between the Northern and Southern lineage.

We note that there are dangers with testing a topology where the ancient individual is not a sister group to the outgroup, since *post-mortem* damage and sequence errors can lead to attraction to the outgroup. However, in this case attraction to the outgroup Han would lead us to reject the tree-like topology, which we do not.

15.5 Comparison of the Anzick-1 individual with Late Pleistocene modern humans from Eurasia

We also computed outgroup f_3-statistics for three previously analysed ancient modern humans: the ~4,000 year old Saqqaq Paleo-Eskimo from Greenland[64], a 40,000-year-old human from Tianyuan in China for, which chromosome 21 was sequenced to low coverage ("Tianyuan"[80]), and a 24,000 year old human from Mal'ta ("MA-1"[29]). For the latter two we used African ascertained SNPs, but for the more recent Saqqaq individual we used the larger SNP array data set that included a large representation of Siberian and New World groups. For the latter two comparisons we sampled a single randomly chosen sequence to call a haploid genotype (both for the Mal'ta, Tianyuan and Anzick-1 individuals), but for the comparison between the Saqqaq- and Anzick-1 individuals that are both sequenced to about 14× coverage we required that the allele was observed in at least 3 sequences.

We find that the Anzick-1 individual is considerably closer to Native American populations than the Saqqaq individual, which in turn is most similar to Greenland Inuit populations and Siberian populations close to the Bering Strait (Chukchi, Koryaks, Yukaghir), in agreement with previous findings (Extended Data Figure 5b and 5c). Compared to MA-1, the Anzick-1 individual is closer to East Asian and Native American populations, while MA-1 is closer to Western Eurasian populations (Extended Data Figure 5d and 5e). This is consistent with a model where the Native American lineage absorbed gene flow from an East Asian lineage as well as a lineage related to the MA-1 individual. Correspondingly, we also find that the Anzick-1 individual is closer to Native American and East Asian populations than the

Tianyuan individual, which at 40,000-years-old seems equally related to a geographically wide range of Eurasian populations in this analysis (Extended Data Figure 5f and 5g).

15.6 The Anzick-1 individual shows the same relative affinity to Western and Eastern Eurasians as present-day First American populations

A previous study has demonstrated that all modern-day Native Americans are likely descendants of an ancient admixture event between a population ancestrally related to modern-day Western Eurasians and the ancestors of East Asians[29]. To assess whether the Anzick-1 individual shows the same relative affinity to Western and Eastern Eurasians, we computed the statistic *D*(Yoruba, Anzick-1; Sardinian, Han) and compared this statistic to that obtained for 52 Native American populations in place of the Anzick-1 individual (*D*[Yoruba, Native American; Sardinian, Han]). We find that the observed statistic for the Anzick-1 individual (D = 0.105±0.006) is similar to that observed in modern-day Native American populations of entirely 'First American' ancestry. This is consistent with the notion that the admixture event occurred prior to the first migration into Beringia and the Americas.

16. ABBA-BABA tests based on sequencing data

16.1 Notation and brief description of the ABBA-BABA test

To investigate the relationship between the Anzick-1 sample and a number of modern populations, we applied an ABBA-BABA test, equivalent to the *D*-statistic based test performed in Green *et al.*[81] and Reich *et al.*[48] to sequencing data from a single genome from each of the populations of interest. If we let H1, H2 and H3 denote 3 human populations including Anzick-1, we used this test to test if the data are consistent with the null hypothesis that tree (((H1, H2), H3), chimpanzee) is correct and there has been no gene flow between H3 and neither H1 nor H2. There are several different definitions of the *D*-statistic in the literature. We used the definition from Durand *et al.*[82]:

$$D = (\text{nABBA-nBABA})/ (\text{nABBA+nBABA})$$

where nABBA is the number of sites where H1 has the same allele as the chimpanzee and H2 and H3 have a different allele (ABBA sites) and nBABA is the number of sites where H2 has the same allele as the chimpanzee and H1 and H3 have a different allele (BABA sites). Under the null hypothesis D=0 and a test statistic that differs significantly from 0 is evidence either of the tree being incorrect or of gene flow (assuming no contamination or differential error

rates). Following Green *et al.*[81] and Reich *et al.*[48], we assessed significance of the deviation from 0 using a Z-score based on blocked jack knife estimates of the standard deviation of the *D*-statistics assuming a block size of 5 Mb. This Z score is based on the assumption that the *D*-statistic (under the null hypothesis) is normally distributed with mean 0 and a standard deviation equal to a standard deviation estimate achieved using the "delete-m Jackknife for unequal m" procedure described in Busing *et al.*[83]

16.2 Data

In the test we included the Anzick-1 genome, along with the genomes listed in Extended Data Table 2.

For the chimpanzee outgroup we used the multiway alignment that includes both chimpanzee and human (pantro2 from the hg19 multiz46).

16.3 Data filtering

As in Orlando *et al.*[49] the data were filtered as follows before we calculated the *D*-statistic:

First all reads with mapping quality below 30 were removed. Subsequently, bases of low quality were removed by dividing all bases into 8 base categories: A, C, G, T on the plus strand and A, C, G, T on the minus strand and then discarding the 50% of the bases with the lowest quality score from each of the 8 categories. More specifically within each base category:

- We found the highest base quality score, Q, for which less than half of the bases in the base category had a quality score smaller than Q.
- We removed all bases with quality score smaller than Q.
- We randomly sampled and removed bases with quality score equal to Q until 50% of the bases from the base category had been removed in total.

After filtering, a single base was sampled for each site for each individual. When the sampled bases included a transition we discarded the site, because the ancient Anzick-1 genome has *post mortem* damage, which strongly increases errors for these sites.

16.4 Test results

We calculated D-statistics and Z-scores for all H1, H2 and H3 configurations with H3=Anzick-1, H2=HGDP00998 (Karitiana) and H1 being one of the rest of the genomes from Extended Data Table 2. Following Green *et al.*[81] and Reich *et al.*[48] we consider absolute

Z-values higher than 3.0 to indicate significant deviations from the null hypothesis, all but one of the genomes got Z-values that indicate significant deviations from the null hypothesis when used as H1. The only exception was the other Karitiana genome (BI16). Note however, that when masking away European ancestry tracks from the Maya genome (HGDP00877) this genome did not lead to a significant Z-value either. Hence with masking, all non-Native American genomes give rise to Z-values that indicate significant deviations from the null hypothesis, and the two Native American samples do not.

17. Test for Anzick-1 ancestry

To assess the relationship between the Anzick-1 individual and other sequenced genomes, we developed a new method for estimating divergence times between pairs of populations using single diploid genomes. This method differs from other methods, such as ML estimation of FST (e.g., [84,85]), in several different ways. Most importantly, by fully parameterizing the ancestral frequency spectrum it makes no assumptions about population genetic processes in the ancestral population. The underlying model, for our new method, has five parameters, the probability of pairwise coalescence before divergence in the first population c_1, the probability of pairwise coalescence before divergence in the second population c_2, and three parameters relating to the allelic configuration in the ancestral populations at the time of divergence ($k_{1,3}$, $k_{2,2}$ and $k_{4,0}$). $k_{1,3}$ is the probability that a variable site had a sample configuration containing one copy of one allele and three copies of another allele in a sample of a total of four gene copies. $k_{2,2}$ and $k_{4,0}$ are similarly defined as the probabilities of a sample configuration of two copies of each of the two alleles, and four copies of just one allele, respectively. Notice that we do not distinguish between ancestral and derived alleles. As we will show, the sample configuration of the two genomes from divergent populations can be specified entirely in terms of these parameters. The different sample configurations of SNPs in the sampled genomes are: (0, 0): invariable, (1, 1): heterozygous in both, (1, 0): heterozygous in the first population and invariable in the second, (0, 1): heterozygous in the second population and invariable in the first, and (0, 2): both homozygous but for different alleles. Assuming no new mutations happened since the time of divergence, the probabilities of these patterns, in terms of the previously defined parameters are

(0, 0): $p_{0,0} = k_{4,0} + c_1c_2(k_{2,2}/3 + k_{1,3}/2) + [(1- c_1)c_2 + c_1(1- c_2)]k_{1,3}/4$

(1, 1): $p_{1,1} = 2(1- c_1)(1- c_2)k_{2,2}/3$

(1, 0): $p_{1,0} = (1- c_1)c_2(k_{1,3}/2 + 2k_{2,2}/3) + (1- c_1)(1- c_2)\, k_{1,3}/2$

(0, 1): $p_{1,0} = (1- c_2)c_1(k_{1,3}/2 + 2k_{2,2}/3) + (1- c_1)(1- c_2)\, k_{1,3}/2$

(0, 2): $p_{0,2} = k_{2,2}(2c_1c_2/3 + [(1- c_1)c_2 + c_1(1- c_2)]/3 + (1- c_1)(1- c_2)/3) + (c_1 + c_2)k_{1,3}/4$

The likelihood function for the parameters Θ = (c1, c2, $k_{1,3}$, $k_{2,2}$, $k_{4,0}$) is then given by $L(\Theta) = \prod_{i \in S} p_i^{n_i}$, where n_i is the number of observed configurations of type i and S = {(0, 0), (1, 1), (1, 0), (0,1), (0, 2)}, which easily can be optimized numerically to obtain a maximum likelihood estimate of Θ.

Notice that this parameterization does not make any assumptions about the underlying demography or about the ancestral population sizes. Inferences under this model are therefore not affected by changes in population sizes in the sampled populations or the ancestral population. However, under specific assumptions regarding population sizes, the parameters c_1 and c_2 can be converted into estimates of divergence times. For example, under the assumption of a constant population size, the divergence time in number of generations for the first population is found by inverting the expression $c_1 = 1- e^{-t/2N}$, i.e, $t = 2N\log(1 - c_1)$, where N is the population size of population 1. Similar interpretations of c_1 and c_2 in terms of divergence times can be found for more complicated demographic models.

The hypothesis that the Anzick-1 individual is a member of a population directly ancestral to a modern population, corresponds to the hypothesis H_0: $c_1 = 0$, if population 1 is Anzick-1 and population 2 is some modern population. This hypothesis can be tested using a likelihood ratio test.

We employ this framework using two modifications to increase robustness. First, we only analyse SNPs that are variable in African populations, as determined by identifying sites that are polymorphic in the set of five African individuals used in this study[86]. We do this to ensure that all SNPs analysed are the product of mutations occurring before divergence of the two populations analysed, as it is an assumption of the model that there are no novel mutations occurring after the time of divergence. This approach is valid as long as the ascertainment of SNPs is carried out in a group that is strictly an outgroup to the two analysed individuals.

The second modification we use is to ignore C/T and G/A polymorphisms. We do that to avoid the effect of damage errors, typical of aDNA, in the Anzick-1 sequence. The parameter estimates for the different populations are in Extended Data Table 4. Note that Extended Data Table 4 only includes estimates for which the Mayan individual is masked for European ancestry. When the Mayan is not masked, using all sites yields a 0.0000 drift units on the

Mayan branch and 0.0069 drift units on the Anzick-1 branch, contrasting the results for the masked Mayan. These parameters were estimated when site pattern counts were n(AAAa) = 150180, n(AaAA) = 152920, n(AAaa) = 49206, n(AaAa) = 97680, and n(AAAA) = 1315136. Additionally, when conditioning only on sites with a C/T or G/A polymorphism, we obtain similar drift estimates for the comparison with the unmasked Mayan of 0.0000 drift units on the Mayan branch and 0.0050 drift units on the Anzick-1 branch. These parameters were estimated when the site pattern counts were n(AAAa) = 46919, n(AaAA) = 48185, n(AAaa) = 15349, n(AaAa) = 31160, and n(AAAA) = 413551.

18. Maximum Likelihood trees of Anzick-1 and other whole-genome sequences

We wanted to assess the relationship of the Anzick-1 individual with respect to a worldwide set of high-coverage whole-genome modern humans sequences[86] and whole-genome ancient DNA sequences from the Saqqaq[64] and Aboriginal Australian[50] genomes. Our analysis utilized *TreeMix*[87] to generate maximum likelihood trees relating these genomes. We removed all sites for which either a C and T or G and A were observed to eliminate any biases due to *post-mortem* deamination in the ancient Anzick-1 and Aboriginal Australian samples. In addition, we retained only sites that were polymorphic in the set of populations and that had data for each population (i.e., no population can have all missing data at the site). We ran *TreeMix* using the San population as outgroup, with blocks of 10 SNPs to account for linkage disequilibrium among sites. We also performed a global rearrangement of the tree after the last step of the graph fitting, and we fit graphs with zero admixture events.

Figure 4c shows a maximum likelihood tree for zero migration events applied to the whole-genome sequencing data (with the Mayan masked). As expected, with zero migration events, the African populations are the most diverged, followed by the European and Central/South Asians splitting off, then the Oceania populations, and then the East Asians, with Anzick-1 being an outgroup to the Mayan and Karitiana individuals. In addition, Figure 4 shows that there is no genetic drift along the Anzick-1 branch, suggesting that the Anzick-1 sample is directly ancestral to the Mayan and Karitiana populations (also reflected in Extended Data Table 4).

19. Outgroup f_3 statistics of Anzick-1 and other whole-genome sequences

We wanted to address which populations the Anzick-1 individual shares the greatest genetic drift. To explore this, we performed outgroup f_3 statistics (see analogous section SI16 on SNPs above and Raghavan *et al.*[29]). Here we use the Yoruban individual as the outgroup population, and the outgroup f_3 statistic between a pair of focal populations will measure the amount of genetic drift on the branch leading from the Yoruban population to the divergence of the focal population pair. For the Anzick-1 and Aboriginal Australian ancient samples, we set the genotypes at a site as missing if a C and T or G and A were observed at the site. This was to mitigate against any biases due to mutations from *post-mortem* deamination.

Extended Data Figure 6 shows results for pairwise outgroup f_3 statistics, comparing affinities of populations either to Saqqaq, Han, or French and comparing affinities of populations either to Anzick-1, Karitiana, or Mayan. With respect to the Saqqaq population, Anzick-1 is always substantially closer to the Mayan or Karitiana populations than to Saqqaq. The Mayan and Karitiana populations are always substantially closer to the Anzick-1 individual than to Saqqaq. Similar comparisons with the Han and French populations show an affinity of the Anzick-1 individual and the Mayan and Karitiana populations.

We also wanted to explore if the Anzick-1 individual has similar affinity to the MA-1 ancient sample[29] from the Mal'ta site as the Karitiana and Mayan individuals, with the Mayan sample being masked for European ancestry. We did not call genotypes for the MA-1 sample due to its low coverage, and instead sampled a single read from each site that overlapped with the polymorphic sites in the whole-genome data. As with the Anzick-1 and Aboriginal Australian ancient samples, for the MA-1 ancient sample we set the genotype at a site as missing if a C and T or G and A were observed at the site to guard against mutations due to *post-mortem* deamination. Extended Data Figure 6 displays pairwise outgroup f_3 statistics, comparing the two modern Native American individuals and the ancient Anzick-1 individual to the ancient MA-1 individual. For all comparisons, the Mayan, Karitiana, and Anzick-1 individuals have similar affinity to the MA-1 individual, implying a similar genetic relationship with the MA-1 individual. Further, the results from Extended Data Figure 6 mirror those observed from SNP chip data in Extended Data Figure 5.

20. References

1. Stanford, D. J. & Bradley, B. A. *Across Atlantic Ice: The Origin of America's Clovis Culture*. (University of California Press, 2012).
2. Lahren, L. A. *Homeland*. (Cayuse Press, 2006).
3. Waters, M. R. & Stafford, T. W. Redefining the age of Clovis: implications for the peopling of the Americas. *Science* **315,** 1122–1126 (2007).
4. Owsley, D. W. & Hunt, D. Clovis and early Archaic crania from the Anzick site (24PA506), Park County, Montana. *Plains Anthropologist* **46,** 115–124 (2001).
5. Goebel, T., Waters, M. R. & O'Rourke, D. H. The late Pleistocene dispersal of modern humans in the Americas. *Science* **319,** 1497–1502 (2008).
6. Morrow, J. E. & Fiedel, S. J. in *Paleoindian Archaeology: A Hemispheric Perspective* (Morrow, J. E. & Fiedel, S. J.) 123–138 (University Press of Florida, 2006).
7. Wilke, P. J., Flenniken, J. J. & Ozbun, T. L. Clovis technology at the Anzick site, Montana. *J Cal Great Basin Anthropol* **13,** 242–272 (1991).
8. Lahren, L. & Bonnichsen, R. Bone Foreshafts from a Clovis Burial in Southwestern Montana. *Science* **186,** 147–150 (1974).
9. Jones, S. & Bonnichsen, R. The Anzick Clovis burial. *Current Research in the Pleistocene* **11,** 42–44 (1994).
10. Santos, G. M., Southon, J. R., Druffel-Rodriguez, K. C., griffin, S. & Mazon, M. Magnesium Perchlorate as an Alternative Water Trap in AMS Graphite Sample Preparation: A Report on Sample Preparation at KCCAMS at the University of California, Irvine. *Radiocarbon* **46,** 165–173 (2004).
11. Scott, E. M., Cook, G. T. & Naysmith, P. A Report on Phase 2 of the Fifth International Radiocarbon Intercomparison (VIRI). *Radiocarbon* **52,** 846–858 (2010).
12. Stafford, T. W., Hare, P. E., Currie, L., Jull, A. J. T. & Donahue, D. J. Study of bone radiocarbon dating accuracy at the University of Arizona NSF accelerator facility for radioisotope analysis. *Radiocarbon* **29,** 24–44 (1987).
13. Yang, D., Eng, B. & Waye, J. Improved DNA extraction from ancient bones using silica-based spin columns. *Am J Phys Anthropol* (1998).
14. Jenkins, D. L. *et al.* Clovis age Western Stemmed projectile points and human coprolites at the Paisley Caves. *Science* **337,** 223–228 (2012).
15. Gonçalves, V. F. *et al.* Identification of Polynesian mtDNA haplogroups in remains of Botocudo Amerindians from Brazil. *Proc Natl Acad Sci U S A* (2013). doi:10.1073/pnas.1217905110
16. Malmström, H. *et al.* Ancient DNA reveals lack of continuity between neolithic hunter-gatherers and contemporary Scandinavians. *Curr Biol* **19,** 1758–1762 (2009).
17. van Oven, M. & Kayser, M. Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation. *Hum Mutat* **30,** E386–94 (2009).
18. Brace, S. *et al.* Population history of the Hispaniolan hutia Plagiodontia aedium (Rodentia: Capromyidae): testing the model of ancient differentiation on a geotectonically complex Caribbean island. *Mol Ecol* **21,** 2239–2253 (2012).
19. Meiri, M. *et al.* Late-glacial recolonization and phylogeography of European red deer (Cervus elaphus L.). *Mol Ecol* **22,** 4711–4722 (2013).
20. Meiri, M. *et al.* Faunal record identifies Bering isthmus conditions as constraint to end-Pleistocene migration to the New World. *Proc R Soc B* **281,** 20132167 (2014).
21. Rohland, N. & Hofreiter, M. Ancient DNA extraction from bones and teeth. *Nat Protoc* **2,** 1756–1762 (2007).
22. Orlando, L. *et al.* Revising the recent evolutionary history of equids using ancient DNA. *Proc Natl Acad Sci U S A* **106,** 21754–21759 (2009).

23. Meyer, M. & Kircher, M. Illumina sequencing library preparation for highly multiplexed target capture and sequencing. *Cold Spring Harb Protoc* **2010,** pdb.prot5448 (2010).
24. Lindgreen, S. AdapterRemoval: Easy Cleaning of Next Generation Sequencing Reads. *BMC Res Notes* **5,** 337 (2012).
25. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25,** 1754–1760 (2009).
26. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25,** 2078–2079 (2009).
27. Schubert, M. *et al.* Improving ancient DNA read mapping against modern reference genomes. *BMC Genomics* **13,** 178 (2012).
28. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26,** 841–842 (2010).
29. Raghavan, M. *et al.* Upper Palaeolithic Siberian genome reveals dual ancestry of Native Americans. *Nature* (2013). doi:10.1038/nature12736
30. DePristo, M. A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* **43,** 491–498 (2011).
31. Wigginton, J. E., Cutler, D. J. & Abecasis, G. R. A note on exact tests of Hardy-Weinberg equilibrium. *Am J Hum Genet* **76,** 887–893 (2005).
32. Maples, B. K., Gravel, S., Kenny, E. E. & Bustamante, C. D. RFMix: A Discriminative Modeling Approach for Rapid and Robust Local-Ancestry Inference. *Am J Hum Genet* **93,** 278–288 (2013).
33. International HapMap 3 Consortium *et al.* Integrating common and rare genetic variation in diverse human populations. *Nature* **467,** 52–58 (2010).
34. Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res* **19,** 1655–1664 (2009).
35. Allentoft, M. E. *et al.* The half-life of DNA in bone: measuring decay kinetics in 158 dated fossils. *Proc R Soc B* **279,** 4724–4733 (2012).
36. Deagle, B. E., Eveson, J. P. & Jarman, S. N. Quantification of damage in DNA recovered from highly degraded samples--a case study on DNA in faeces. *Front Zool* **3,** 11 (2006).
37. Hijmans, R. J., Cameron, S. E., Parra, J. L., Jones, P. G. & Jarvis, A. Very high resolution interpolated climate surfaces for global land areas. *Int. J. Climatol.* **25,** 1965–1978 (2005).
38. Braconnot, P. *et al.* Results of PMIP2 coupled simulations of the Mid-Holocene and Last Glacial Maximum – Part 1: experiments and large-scale features. *Clim. Past* **3,** 261–277 (2007).
39. Bintanja, R., van de Wal, R. S. W. & Oerlemans, J. Modelled atmospheric temperatures and global sea levels over the past million years. *Nature* **437,** 125–128 (2005).
40. Briggs, A. W. *et al.* Patterns of damage in genomic DNA sequences from a Neandertal. *Proc Natl Acad Sci USA* **104,** 14616–14621 (2007).
41. Jónsson, H., Ginolhac, A., Schubert, M., Johnson, P. & Orlando, L. mapDamage2.0: fast approximate Bayesian estimates of ancient DNA damage parameters. *Bioinformatics* (2013). doi:10.1093/bioinformatics/btt193
42. Krause, J. *et al.* A Complete mtDNA Genome of an Early Modern Human from Kostenki, Russia. *Curr Biol* **20,** 231–236 (2010).
43. Milne, I. *et al.* Using Tablet for visual exploration of second-generation sequencing data. *Brief. Bioinformatics* **14,** 193–202 (2013).
44. Li, M., Schroeder, R., Ko, A. & Stoneking, M. Fidelity of capture-enrichment for mtDNA genome sequencing: influence of NUMTs. *Nucleic Acids Res* **40,** e137–e137 (2012).

45. Andrews, R. M. *et al.* Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA. *Nat Genet* **23,** 147 (1999).
46. Altschul, S., Gish, W., Miller, W., Myers, E. & Lipman, D. Basic local alignment search tool. *J Mol Biol* **215,** 403–410 (1990).
47. Fu, Q. *et al.* A revised timescale for human evolution based on ancient mitochondrial genomes. *Curr Biol* **23,** 553–559 (2013).
48. Reich, D. *et al.* Genetic history of an archaic hominin group from Denisova Cave in Siberia. *Nature* **468,** 1053–1060 (2010).
49. Orlando, L., Ginolhac, A., Zhang, G. & Froese, D. Recalibrating Equus evolution using the genome sequence of an early Middle Pleistocene horse. *Nature* (2013). doi:doi:10.1038/nature12323
50. Rasmussen, M. *et al.* An Aboriginal Australian genome reveals separate human dispersals into Asia. *Science* **334,** 94–98 (2011).
51. Frazer, K. A. *et al.* A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449,** 851–861 (2007).
52. Tamm, E. *et al.* Beringian standstill and spread of Native American founders. *PLoS ONE* **2,** e829 (2007).
53. Perego, U. A. *et al.* Distinctive Paleo-Indian migration routes from Beringia marked by two rare mtDNA haplogroups. *Curr Biol* **19,** 1–8 (2009).
54. Achilli, A. *et al.* The phylogeny of the four pan-American MtDNA haplogroups: implications for evolutionary and disease studies. *PLoS ONE* **3,** e1764 (2008).
55. Cui, Y. *et al.* Ancient DNA analysis of mid-holocene individuals from the Northwest Coast of North America reveals different evolutionary paths for mitogenomes. *PLoS ONE* **8,** e66948 (2013).
56. Yao, A., Kong, Q., Bandelt, H., Kivisild, T. & Zhang, Y.-P. Phylogeographic Differentiation of Mitochondrial DNA in Han Chinese. *Am J Hum Genet* **70,** 635–651 (2002).
57. Behar, D. M. *et al.* A 'Copernican' reassessment of the human mitochondrial DNA tree from its root. *Am J Hum Genet* **90,** 675–684 (2012).
58. Rickards, O., Martínez-Labarga, C., Lum, J. K., De Stefano, G. F. & Cann, R. L. mtDNA history of the Cayapa Amerinds of Ecuador: detection of additional founding lineages for the Native American populations. *Am J Hum Genet* **65,** 519–530 (1999).
59. Kemp, B. M. *et al.* Genetic analysis of early holocene skeletal remains from Alaska and its implications for the settlement of the Americas. *Am J Phys Anthropol* **132,** 605–621 (2007).
60. García-Bour, J. *et al.* Early population differentiation in extinct aborigines from Tierra del Fuego-Patagonia: ancient mtDNA sequences and Y-chromosome STR characterization. *Am J Phys Anthropol* **123,** 361–370 (2004).
61. Bolnick, D. A. & Smith, D. G. JSTOR: American Antiquity, Vol. 72, No. 4 (Oct., 2007), pp. 627-644. *American antiquity* (2007).
62. Cann, H. M. *et al.* A human genome diversity cell line panel. *Science* **296,** 261–262 (2002).
63. Poznik, G. D. *et al.* Sequencing Y chromosomes resolves discrepancy in time to common ancestor of males versus females. *Science* **341,** 562–565 (2013).
64. Rasmussen, M. *et al.* Ancient human genome sequence of an extinct Palaeo-Eskimo. *Nature* **463,** 757–762 (2010).
65. 1000 Genomes Project Consortium *et al.* An integrated map of genetic variation from 1,092 human genomes. *Nature* **491,** 56–65 (2012).
66. Tamura, K. *et al.* MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol Biol Evol* **28,** 2731–2739 (2011).

67. Dulik, M. C. *et al.* Y-chromosome analysis reveals genetic divergence and new founding native lineages in Athapaskan- and Eskimoan-speaking populations. *Proc Natl Acad Sci U S A* **109,** 8471–8476 (2012).
68. Regueiro, M., Alvarez, J., Rowold, D. & Herrera, R. J. On the origins, rapid expansion and genetic diversity of Native Americans from hunting-gatherers to agriculturalists. *Am J Phys Anthropol* **150,** 333–348 (2013).
69. Pritchard, J. K., Stephens, M. & Donnelly, P. Inference of population structure using multilocus genotype data. *Genetics* **155,** 945–959 (2000).
70. Hancock, A. M. *et al.* Adaptations to climate-mediated selective pressures in humans. *PLoS Genet* **7,** e1001375 (2011).
71. Li, J. Z. *et al.* Worldwide human relationships inferred from genome-wide patterns of variation. *Science* **319,** 1100–1104 (2008).
72. Reich, D. *et al.* Reconstructing Native American population history. *Nature* **488,** 370–374 (2012).
73. Surakka, I. *et al.* Founder population-specific HapMap panel increases power in GWA studies through improved imputation accuracy and CNV tagging. *Genome Res* **20,** 1344–1351 (2010).
74. Behar, D. M. *et al.* The genome-wide structure of the Jewish people. *Nature* **466,** 238–242 (2010).
75. Fedorova, S. A. *et al.* Autosomal and uniparental portraits of the native populations of Sakha (Yakutia): implications for the peopling of Northeast Eurasia. *BMC Evol Biol* **13,** 127 (2013).
76. Metspalu, M. *et al.* Shared and unique components of human population structure and genome-wide signals of positive selection in South Asia. *Am J Hum Genet* **89,** 731–744 (2011).
77. Yunusbayev, B. *et al.* The Caucasus as an asymmetric semipermeable barrier to ancient human migrations. *Mol Biol Evol* **29,** 359–365 (2012).
78. Patterson, N. *et al.* Ancient admixture in human history. *Genetics* **192,** 1065–1093 (2012).
79. Reich, D., Thangaraj, K., Patterson, N., Price, A. L. & Singh, L. Reconstructing Indian population history. *Nature* **461,** 489–494 (2009).
80. Fu, Q. *et al.* DNA analysis of an early modern human from Tianyuan Cave, China. *Proc Natl Acad Sci U S A* **110,** 2223–2227 (2013).
81. Green, R. E. *et al.* A Draft Sequence of the Neandertal Genome. *Science* **328,** 710–722 (2010).
82. Durand, E. Y., Patterson, N., Reich, D. & Slatkin, M. Testing for ancient admixture between closely related populations. *Mol Biol Evol* **28,** 2239–2252 (2011).
83. Busing, F. M., Meijer, E. & Leeden, R. V. D. Delete-m Jackknife for Unequal m. *Statistics and Computing* **9,** 3–8 (1999).
84. Nicholson, G. *et al.* Assessing population differentiation and isolation from single-nucleotide polymorphism data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **64,** 695–715 (2002).
85. Balding, D. J. Likelihood-based inference for genetic correlation coefficients. *Theor Popul Biol* **63,** 221–230 (2003).
86. Meyer, M. *et al.* A High-Coverage Genome Sequence from an Archaic Denisovan Individual. *Science* (2012). doi:10.1126/science.1224344
87. Pickrell, J. K. & Pritchard, J. K. Inference of population splits and mixtures from genome-wide allele frequency data. *PLoS Genet* **8,** e1002967 (2012).