doi:10.1038/nature25173

Terminal Pleistocene Alaskan genome reveals first founding population of Native Americans

Supplementary Information

1. Archaeological context

Two infants (USR Individuals 1 and 2) were recovered from a burial pit at the Upward Sun River site (USR), located in the Tanana River valley in central Alaska (49XBD-298) (Figure S1). The two individuals have been named Xach'itee'aanenh t'eede gaay or sunrise child-girl (USR1) and Yełkaanenh t'eede gaay or (dawn twilight child-girl (USR2) by the local Native community. The archaeological and geological context of these finds has been previously reported [1,2]. The infants were discovered in a pit within a massive aeolian silt (Unit IV) at ~160-170 cm below surface. Post-depositional disturbance is interpreted as minimal given undisturbed Ab horizons across the site with smooth and very abrupt horizon boundaries and a thin vertical distribution of cultural materials [2].

The two infants are associated with a burial pit (F2011-13) within a residential feature in Component 3 at USR dating to the terminal Pleistocene – Holocene transition [2]. The cremated child (USR Individual 3) was found within a hearth (F2010-5) associated with burned bone extending to about 43 cm below the occupation surface. This Component 3 surface is composed of numerous lithic and charcoal fragments in a thin, unimodal vertical distribution. Four charcoal samples from the base and top of the pit hearth as well as from the burial pit yielded statistically similar radiocarbon dates, with a mean pooled radiocarbon age of 9970+/-30 BP (11,600-11,270 cal BP), consistent with other Component 3 hearth dates [1–3]. The burial pit was excavated through an earlier hearth. After burial of Individuals 1 and 2, the pit was partially backfilled and saw continued use as a hearth. Individual 3 was cremated within this hearth and then the residential feature was abandoned [2]. Faunal remains associated with both the backfilled burial pit and hearth indicates use of salmon, genetically and isotopically confirmed as anadromous *Onchorynchus keta* (chum salmon), representing the earliest human use of salmon in the New World [4], as well as *Urocitellus parryi* (ground squirrel), *Lepus americanus* (hare), and other mammalian and avian taxa, including an unidentified very large mammal [2].

Based on the dimensions of diagnostic skeletal elements and deciduous crown development [5,6], Individual 1 died at 6-12 weeks post-natal and Individual 2 died at >30 prenatal weeks. Morphological analyses suggest that both individuals are female [7]. Genetic data presented in this paper confirms this assessment (Section 4.6). Results from dental morphological trait analysis indicate Individual 1 is consistent with a Native American population affinity [2,8].

Petrous specimens were selected from each individual for aDNA analyses given their mass, high density and reported consistent high endogenous DNA yield [9,10]. Specimen #58-311 (USR1) weighed 3.4 g and specimen #58-80 (USR2) weighed 1.2 g. Tackney et al [11] extracted and sequenced the mitochondrial genomes from both individuals, providing the first sample of ancient Beringian genetic diversity. Individual 1 was within the C1b lineage and Individual 2 was placed at the root of the B2 lineage. Both

mtDNA haplogroups are not currently found near Beringia among recent Native American populations. Subsamples from these two petrous specimens were used in this current study for nuclear DNA extraction and genomic analyses.

Cultural affiliation for all three individuals, and USR Component 3, is established through various lines of evidence. The USR projectile points recovered within the burial are most consistent with Eastern Beringian Denali complex foliate forms [12–14], and are dissimilar to contemporaneous Mesa and Sluiceway complex forms, the latter found in the Brooks Range [15,16]. The contemporaneous Northern Fluted Point complex has characteristic concave-based fluted points but lacks foliate forms [17]. During the period of Component 3 occupation, Denali complex sites are common throughout interior Alaska and the Yukon Territory (Figure S1), including numerous sites within 50 km of USR (e.g., Dry Creek, Mead, Swan Point and Gerstle River, Alaska) [18]. The Denali complex dates between 12,000 and 6,000 cal BP and is widespread throughout interior eastern Beringia as far west as Yukon Territory [14,18–21], and possibly northern Alberta [22]. Thus, the individuals associated with USR Component 3 are part of a widespread material cultural tradition extending through much of Northwestern North America during the late Pleistocene and early Holocene.

The American Paleoarctic tradition / Denali complex was entirely replaced in both interior and Alaskan western and northern coastal areas by the Northern Archaic tradition at ca. 6,000 cal BP [14,18,23]. Some technologies continued through this transition, including Campus-type wedge shaped microblade cores and edge-ground convex based projectile points, while new technological and typological materials included side-notched bifaces, semi-lunar bifaces, and side-notched cobbles. The Denali – Northern Archaic transition is not well understood, and may reflect migration of a new population into the region, diffusion of technology with population continuity, or mixture of two populations (see review in [24]).

2. Laboratory procedures

All ancient DNA work, prior to library amplification, was conducted in dedicated clean lab facilities at the Centre for GeoGenetics, Natural History Museum, University of Copenhagen, according to strict ancient DNA (aDNA) guidelines [25,26]. Bone powder was prepared from remains of pars petrosa obtained from individuals 1 and 2 (USR1 and USR2) from the Upward Sun River site [1,2]. DNA extraction and preparation of libraries for Illumina sequencing followed previously published protocols [27]. Briefly, bone powder was digested in 0.5 M EDTA supplemented with 10 mM TE-buffer pH=8, Proteinase K (0.14-0.22 mg/ml) and 0.5% w/v N-laurylsacosine at 37°C overnight. Sample supernatant was incubated for 1 hour with size-selected silica powder (50 ul) [28] in 10 volumes of improved binding buffer [27]. Upon binding, silica pellets were washed twice in 1 ml 80% ethanol, and eluted in TE-buffer with 0.05% Tween-20.

Libraries were prepared from double-stranded DNA fragments following the methods described elsewhere [29]. For some libraries the DNA extracts were pre-treated with 5 ul USER enzyme (#M5505, NEB) for 3 hours at 37°C. Extracts were end-repaired using NEBnext end-repair module (#E6050, NEB). Adaptors (12.5 pmol/rxn), prepared according to Meyer and Kircher [29] were ligated onto library DNA using the NEBnext Quick Ligation Module (#E6056 NEB). After end-repair and adaptor ligation DNA was recovered using Qiagen MinElute columns. Fill-in reactions were performed using Bst large fragment (#M0275 NEB) in the accompanying buffer supplemented with dNTPs. Index PCR reactions were performed with KapaU+ DNA polymerase (KapaBiosystems) and 400 nM primer introducing hexamer sequencing index motifs in each as well as the required P5 and P7 sequences. The number of cycles for index PCRs was determined from qPCR analysis. Purified libraries were sequenced at the National High-throughput Sequencing Centre, University of Copenhagen.

3. Sequence data processing

3.1. Basecalling and adapter removal

Base calls were produced using the Illumina software *CASAVA 1.8.2*. Sequence data were then demultiplexed by requiring a complete match to the 6 and 10 nucleotide index sequences used for library preparation. Dual-indexed paired-end raw reads were trimmed for Illumina adapter sequences, leading Ns (--trimns) and trailing quality 2 runs (--trimqualities --minquality 2) using *AdapterRemoval v1.5.3* [30]. Overlapping pairs were collapsed (--collapse) into single reads requiring a minimum overlap of 11 bases. We only kept collapsed reads with length greater than 30 bases (--minlength 30) for subsequent analyses. Other sequencing data not produced as part of this study (Section 6.2) were processed as described in [31].

3.2. Mapping

Mapping of the trimmed reads was done following the strategy in [31]. Reads were mapped to the human reference genome build 37 using *bwa v0.6.2-r126* [32], following [33]. Filtering and post-processing was carried out using *picard tools* (http://picard.sourceforge.net) and GATK [34]. Finally, the MD tag was recomputed for each read using the *samtools* 0.1.18 [35] *calmd* command. We report mapping statistics in Table S1. Sequencing data not produced as part of this study were processed as described in [31].

3.3. Genotyping

We called diploid genotypes from the mapped reads following [31]. We ran *samtools mpileup* [35] with the *-C* parameter set to 50 and only included sites with a depth of coverage greater than 1/3 the average depth and lower than 2 times the average depth (6 and 34X, respectively, for USR1). We then filtered out variant calls with a phred

posterior probability lower than 30, significant strand and/or end distance bias ($p<1e-4$) or located within 5 bp of each other. Additionally, we discarded heterozygote calls with an allelic ratio lower than 0.2. Diploid calls were then merged with the genotype data from [31] and sites with a significant deviation from the Hardy-Weinberg equilibrium ($p<1e-4$) were filtered out [36].

3.4. Phasing

For the analyses requiring haplotype data (Section 18), the filtered callset was phased using *shapeit2-r727* [37], following the strategy from [31], and assuming an effective population size of 20,000.

4. Ancient DNA sequence data assessment

4.1. Ancient DNA misincorporation patterns

It has been shown that in the absence of cellular repair mechanisms, DNA gradually accumulates chemical changes, including for example, double stranded breaks and cytosine deamination [26]. Such changes have been used as a means of authenticating ancient DNA (aDNA) sequence data [38]. Using *bamdamage* [39], for all extracts we found that the fragment length distribution was biased towards shorter fragments (mean read length ranged from 47.5 to 59.1 bp in non-USER treated extracts). In addition, we found an increased frequency of C to T substitutions towards the 5' ends of the reads (10.9% to 14.2% in the first 5' nucleotide in non-USER treated extracts), as well as the complementary G to A substitutions towards the 3' ends (10.5% to 13.5% in the first 3' nucleotide) (Figure S2, Table S1). These patterns suggest that the extracted DNA is effectively endogenous to the sampled human remains.

4.2. Sequencing error rates

We obtained estimates for the overall and type-specific sequencing error rates using the method described in [40]. These estimates are based on the assumption that any pair of human genomes carries on average the same number of derived alleles compared to an outgroup. Therefore, any excess derived alleles in a given sample should be due to error.

We obtained the number of expected derived alleles by using the high quality sequence data for individual NA12778 from the 1000 Genomes Project. We determined the ancestral allele based on the chimp genome, as included in the *pantro2-hg19* multiZ alignments from the UCSC Human Genome Browser [41]. We then used the implementation available in *ANGSD* [42] to estimate the error rates in our sequence data, relative to the high quality data from the 1000 Genomes project. In agreement with proposed patterns of aDNA degradation [38], C to T and G to A substitutions accounted for the majority of the errors. Note that error is decreased when considering USER-

treated libraries [43]. Therefore, we restricted all subsequent analyses to data coming from USER-treated libraries. This resulted in an overall sequencing error rate of 0.09% for USR1 and 0.11% for USR2. The breakdown of relative error rates by library is presented in Table S1 and Figure S2.

4.3. Genotyping error rates

We estimated overall error rates for the filtered diploid genotype calls using the same approach that was used for estimating sequencing error [40] (Section 4.2). In this case, we used the human reference genome (hg19), which we assume to be devoid of errors, to obtain the number of expected derived alleles compared to the chimp genome. We then obtained the number of observed derived alleles by comparing the two called alleles to the chimp allele, at sites where all three genomes have non-missing data.

We obtained an overall genotyping error rate estimate of 0.0023% for USR1. As expected, this value was lower than the estimated sequencing error rate (Section 4.2). For comparison, we also estimated genotyping error rates for the ancient Anzick1 genome (0.0056%), as well as for the modern Papuan (0.0025%), Han (0.0024%), and Aymara (0.0022%) genomes. In contrast to USR1, the Anzick1 genome [44] was not based on USER-treated libraries, thus yielding a higher estimate. In fact, the estimated genotyping error for USR1 is comparable to those of the other modern genomes. Given USR1's low genotyping error estimate (0.0023%) and how similar that is to those of modern genomes, we consider that genotyping error is unlikely to bias subsequent analyses of this ancient genome.

4.4. mtDNA-based contamination estimates

In order to quantify the level of contamination present in the data, we used the mtDNA-based method described in [45]. This method assumes that contamination and error account for less than 50% of the data and that depth of coverage is high enough to call the real endogenous mtDNA consensus sequence. Sequence reads are then modeled as coming from a mixture of a set of full-length mitochondrial genomes (*311* mitochondrial genomes from [46]) and the 'endogenous' consensus with unknown proportions. Such proportions are in turn, inferred using an MCMC method. Thus, the contamination fraction is the sum of the proportions corresponding to the 'non-endogenous' mitochondrial sequences.

To estimate contamination in our data, we first built an mtDNA consensus sequence using *ANGSD* [42] (-doFasta 2). Reads with a mapping quality lower than 30 and nucleotides with a base quality lower than 20 were discarded and sites with a depth of coverage lower than 5X were set to 'N'. In order to reduce potential noise introduced by reads originated from mitochondrial inserts in the nucleus [47], only the reads that mapped uniquely to the mtDNA were then remapped to the newly built consensus sequence. The consensus sequence was also aligned to a panel of 311 worldwide mtDNA sequences [46]

using mafft [48,49]. Finally, we used both the alignment and the mapped reads for contamination estimation. We obtained contamination estimates for each sequenced library with mtDNA coverage greater than 5X, as well as for the merged data for each sample (Table S1) by running five independent chains for 50,000 iterations. For each run, convergence was confirmed by monitoring the Gelman-Rubin diagnostic [50]. For USR1, we estimated modern contamination to be 0.15% with a 95% credibility interval of (0.03% - 0.43%) and for USR2, we estimated contamination to be 0.05% with a 95% credibility interval of (0.01% - 1.15%). These estimates are comparable to those obtained for other ancient samples, *e.g.* the Anzick-1 genome [44]. Moreover, given their magnitude, we deem modern contamination unlikely to substantially bias the population genetics results presented below.

4.5. mtDNA haplogroup assignment

Mitochondrial DNA from both USR individuals has been reported previously [11] (Section 1). To confirm these results, we built mtDNA consensus sequences for both individuals (Section 4.3) and used the *HAPLOFIND* [51] web service to assign them to a mtDNA haplogroup. USR1 was found to carry the C1b mtDNA haplogroup, while USR2 was assigned to the B2 haplogroup. While both haplogroups have been found in high frequencies in Native American populations [52], their frequencies are low in North American populations. As discussed in [11], these results suggest that genetic variation characteristic of contemporary Native Americans was already present in early Beringian populations (Section 1).

4.6. Sex chromosome karyotype determination of the USR remains

Based on morphological analyses, both non-cremated individuals from the USR site (USR1 and USR2) were previously identified as female [2]. We confirmed these assignments using sequencing data. We computed the proportion of filtered reads that could be mapped to the Y chromosome (R_y), with respect to the number of reads mapped to the X chromosome as detailed in [53]. We found R_y to be ~0.0032 +/- 2.3e-^{13} for USR1 and ~0.005 +/- 5.6e^{-10} for USR2. Based on the calibration panel from [53], a conservative estimate for an XY karyotype is Ry>0.075, and Ry<0.016 for an XX karyotype. Thus, we conclude that the karyotype in both USR individuals is XX, suggesting both individuals were female (Section 1).

4.7. Nuclear contamination estimates

In addition to mtDNA-based contamination estimates, we used DICE [54] to estimate modern human contamination in USR1, based on the nuclear genome. DICE models the observed sequencing data as a joint function of the sequencing error, the contamination rate and the demographic history underlying the sample of interest and the contaminant. Likelihood scores regarding the latter are computed through a diffusion-based approach

and an MCMC algorithm is used to obtain maximum *a posteriori* estimates for said parameters.

We used the two-population model implemented in DICE. To do so, we used the 1000 Genomes Project 'CEU' allele frequencies as a proxy for the frequencies of the putative contaminant and the 'YRI' frequencies to represent the 'anchor' population. Note that we only obtained nuclear contamination estimates on the final USR1 dataset, given that DICE is dependent on high-depth data [54]. We let the MCMC algorithm run for 100,000 steps from which we burnt the first 10,000. We then used the *coda* R package to obtain the following 95% posterior confidence intervals:

Contamination rate: 0.1401 - 0.1414%
Error rate: 0.0461 - 0.0467%
Drift parameter leading to USR1: 0.524 - 0.527
Drift parameter leading to the contaminant (CEU): 0.0316 - 0.0324.

The estimated contamination rate is comparable to that obtained with the mtDNA-based method (Section 4.4). Given the magnitude of these estimates, we consider that modern human contamination is unlikely to introduce any significant bias to nuclear genome-based results.

5. Relatedness analysis of USR children burial

We applied two approaches to identifying the familial relationship between the two USR individuals. However, due to the very low sequencing depth of one of the individuals (USR2) we cannot use standard approaches based on called genotypes, as this would give a biased result. Instead we tried to infer the identity by descent (IBD) sharing between the samples using *NGSrelate* [55], which can be applied directly to genotype likelihoods, thereby taking the uncertainty of the genotypes into account. This method also uses allele frequencies from the sample population. However, since we only have two individuals, these cannot be obtained. Instead we used the Peruvian (PEL) Native American samples from the 1000 genomes data set as a proxy. If the allele frequencies do not match the target population perfectly we expect a bias in the results. Therefore, we also tried a slightly different approach by modeling the IBD sharing tracts across the genome, which potentially can alleviate some of this bias. For this, we used *relate* [56] which can also be applied to genotype likelihoods. This method tries to model the linkage disequilibrium (LD); however, since the Peruvian samples are only a proxy for the LD patterns in the USR samples we instead chose to 'prune' away the LD in our data. Because the Peruvian samples also contain gene flow from Europe we estimated the ancestral allele frequencies based on *ADMIXTURE* results [57].

5.1. Data

Allele frequencies were estimated based on the genotypes from the CEU, PEL, CBH and YRI individuals in the 1000 genomes data (VCF files, release20130502). These were analyzed using *ADMIXTURE* assuming K=4 populations in order to obtain the ancestral allele frequencies of the Peruvian component. SNPs with a minor allele frequency below 5%, and SNPs not located on the autosomes were removed.

We used sequencing data from the 2 USR individuals as well as the ancient Anzick1 sample [44]. Genotype likelihoods were calculated for the sequenced individuals using ANGSD [42], based on the samtools genotype likelihood model. The major and minor alleles were determined from the 1000 genomes data. Reads with a mapping quality below 30 and bases with a base quality below 20 were removed. Additionally, we required a minimum of 2 reads for the genotype likelihood calculations.

For the *relate* based analysis, we also decreased LD by 'pruning' sites with an $r^2>0.1$ using PLINK [58] based on the Peruvian samples in a 500 SNP window. This was done separately for each pair of individuals based only on the sites where genotype likelihoods were available for both individuals.

5.2. Results

The results from *NGSrelate* are provided in Table S2. As can be seen there we estimate that Anzick1 is slightly related to the USR samples. Since these individuals are from different regions (separated by ~4000 km) and time periods (~1300 years) this is obviously not true, but likely caused by the bias of using the wrong population allele frequencies.

The results from *relate* where we try to model the IBD tracts across the genome is shown in Table S2. For USR2 the results are as expected: it is unrelated to the Anzick1 individual. However, this result is based on less than a thousand sites. For USR1, the results remain equally biased. Therefore, we were not able to remove the bias from using the Peruvian allele frequencies as a proxy. However, the results still show that the two USR individuals have a much higher estimated relatedness. Assuming that the estimates are reliable we would conclude that the two individuals are within the range of half siblings. However, since the methods seem to bias the results upwards, it is likely that they are more distantly related, such as first cousins.

6. Reference data

We compared the genomes of the USR individuals to a SNP array dataset and a set of contemporary and ancient genomes.

6.1. Genotype panel

We used a genotype dataset consisting of 2,537 contemporary individuals from 167 ethnic groups (enriched in Native Americans), genotyped over 199,285 SNP sites [31,59–61]. In order to avoid any confounding signal caused by recent admixture, Native American individuals included in this panel were masked for European and African admixture by using the local ancestry assignment approach implemented in *RFmix* [62]. This dataset is a subset of the panel assembled in [31], from which we filtered out all individuals with more than 95% missing data (including masking). Note that we used the masks obtained through 'Method 1', as described in the Supplementary Material of [31]. The Native American populations in the genotype panel were assigned following the language-based classification used in [60], originally based on [63].

6.2. Whole genome sequence dataset

In addition to the genotype dataset described in Section 6.1, we used a set of contemporary and ancient whole genome sequences from [31,44,59,64–69], as assembled in [31] (Table S3). Note that for the Mixe and Surui genomes from [68], we used the genotypes distributed by the Simons Foundation dataset. Average depth of coverage ranged from 19X to 44.2X for modern genomes, from 1X to 14X for ancient human genomes, and from 24.3X to 40.8X for archaic hominin genomes (Table S3). We computed the fraction of heterozygote calls with respect to the total non-missing genotypes for each genome and found Africans to carry the highest proportion of heterozygote calls, followed by Eurasians, Papuans and Native Americans. This pattern is in agreement with previous measurements in related populations, and recapitulate serial founder effects out of Africa, *e.g.* [70]. In agreement with its geographic location, we found USR1 to carry a heterozygote genotype proportion similar to that of other Native Americans. Sequencing data was processed as detailed in [31] and is summarized in Table S3. Note that for the Aleutian individual, we only considered the fraction of the genome that does not correspond to recent European admixture, as identified in [31].

7. Multidimensional scaling

We used multidimensional scaling (MDS) to explore the relationship between the genome of USR1, a set of ancient genomes (Section 6.2) and a genotype panel of worldwide populations (Section 6.1). Given that a subset of the ancient genomes was sequenced to a low depth of coverage (which renders genotype calling unfeasible), we used the random allele sampling strategy implemented in *bammds* [39]. In brief, for each site included in the reference panel, we sampled one random base with mapping quality ≥30 and base quality ≥20 for sequence data and one random allele for genotype data, and computed an identity-by-state (IBS) distance matrix that was then used to obtain an MDS transformation.

We built MDS plots using two different subsets of the genotype panel. First, we considered all non-African individuals. In this projection, USR1 fell adjacent to Native American and Siberian variation (Figure S3a). We then built a second MDS plot to

explore the relationship between USR1, Siberians and Native Americans. In this case, USR1 was placed in an intermediate position relative to the Siberian and Native American clusters (Figure S3b). We note that while Eskimo-Aleut speaking populations are similarly placed in an intermediate position due to more recent Asian gene flow into these populations [60,71] (first dimension), USR1 was also placed between the two major Native American groups (second dimension) that we refer to as the Northern and Southern Native American branches (NNA and SNA) [44]. See Section 10.3 for details on the SNA/NNA classification. Given the geographic origin and age of the USR1 sample (Section 1), these results raise the possibility that the USR1 genome represents a population that branched off the Native American lineage after divergence from an Asian source population, but prior to the diversification of those basal Native American lineages.

8. Admixture

8.1. Methods

We explored the genetic ancestry components carried by USR1 and a set of ancient genomes through a model-based clustering analysis. We used a genotype likelihood-based approach to infer ancestry proportions from sequence data, based on the ancestral components inferred for the modern genotype panel. We first performed a separate *ADMIXTURE* [57] analysis on the genotype panel (Section 6.1) with different numbers of ancestral components ($K=\{2,...,20\}$). For each value of K, we ran 100 replicates with different seed values and selected the run with the highest likelihood. We then computed the genotype likelihoods for each site in the panel, for each of the ancient genomes using the *GATK* model as implemented in *ANGSD* [42]. We only considered reads with mapping quality >=30, bases with base quality >=20 and discarded 5 nucleotides from each end of the reads. Finally, for each value of K, we used the allele frequencies inferred by *ADMIXTURE* and the genotype likelihoods as an input for the expectation maximization algorithm from [72] (implemented in http://www.popgen.dk/software/index.php/FastNgsAdmix) to estimate the ancestry proportions for each ancient genome.

8.2. Admixture results for the reference panel

We first ran ADMIXTURE on the reference panel to estimate the allele frequencies for each ancestral component (Figure S3c). When we assumed low numbers of ancestral components, continental scale genetic differentiation was readily observed. Native American, East Asian, European and Oceanian components arose at $K=2,3,4$ and 5, respectively. In the following, we describe population structure patterns for K values ($K=8,9,10,12,13,15$ and 20) at which 'new' Native American ancestry components were observed.

- *K=8*. Chukchi-Kamchatkan-speaking populations from Northeast Siberia differentiated from other Siberian populations. This component replaced the Siberian component present in Native American populations. This pattern is consistent with recent gene flow between northeastern Siberians and northern Native Americans, which has been previously documented [60,61,71] and that we also identify in Section 10.2.
- *K=9*. Chibchan-Paezan-speaking populations formed their own cluster. However, other Native American populations carry a low amount of the Chibchan-Paezan component.
- *K=10*. An ancestry component mainly associated with Equatorial-Tucanoan speaking populations arose. While this component is present in Central and Northern-Amerind-speaking populations, we found the Brazilian Surui to derive, on average, more than 99% of their ancestry from this component.
- *K=12*. A North Native American (NNA) ancestry component arose in Na Dene and Northern Amerind-speaking populations. Central Amerind-speaking populations carry a small proportion of said component. On the other hand, populations from South America were found to be approximately devoid of such signal.
- *K=13*. An ancestry component characteristic to Central Amerind-speaking populations arose. This component is widely spread through Central and North American populations with the Pima carrying the highest proportion. Of note is the fact that Na Dene- and Chibchan-Paezan-speaking populations carry the lowest proportion of this component.
- *K=15*. A second Equatorial-Tucanoan ancestry component arose. This component has a range similar to that of the Surui. However, it differentiates the Brazilian Karitiana from other nearby populations.
- *K=20*. Greenlandic Inuits formed a cluster separate from other Native American populations. As expected [60,61,71], we found this component to be present in the Siberian Chukchi and Naukan.

8.3. USR1 bears Asian and Native American ancestry

To learn about the ancestry of USR1 at a continental level, we estimated its ancestry proportions when assuming *K=12* admixture components (Figure S3c). We found that USR1 carries ~56.3% of the Native American ancestry components, which was further stratified into similar proportions of the South (~30.9%) and North (~25.4%) Native American components. Meanwhile, the remaining ~43.7% was assigned to Eurasian ancestry components, for which Siberians carry the most similar distribution among the populations in the panel. In addition, we found that other Native American genomes do not carry such Siberian component. While the ancient NNA genome '939' (sequenced from ~6 kya skeletal remains from the Lucy Islands, British Columbia, Canada) [31] and the early diverged Kennewick Man [67] carry the Chukchi-Kamchatkan-specific

admixture component, we did not detect the Siberian or the East Asian components in any of those genomes or in the SNA Anzick1 genome (Figure S3c). Note that the population from which USR1 originated is not represented in the panel. Therefore, USR1 is expected to carry a mixture of the ancestral components that are genetically closer: namely, Native Americans and to a lesser extent Siberians. Together with the MDS projection (Figures S3a and S3b), this result supports the conclusion that the USR1 lineage represents a population that diverged from Native Americans prior to their diversification into the North and South lineages.

8.4. USR1 carries ancestry related to all major Native American groups

To better characterize the genetic relationship between USR1 and contemporary Native American populations, we examined the ancestry proportions in USR1, when assuming *K=20* ancestral populations. At this value of *K*, Native American populations carry different proportions of the seven Native American-specific ancestry components (Figure 1d). We found that USR1 carries 17.38% from the SNA component, 0.45% from the Surui Equatorial-Tucanoan component, 2.75% from the Karitiana Equatorial-Tucanoan component, 1.78% from the Chibchan-Paezan component, 8.43% from the Central Amerind component, 23.91% from the Na Dene component and 4.29% from the Inuit component. While other ancient Native American genomes carry these same components, their distributions are more biased towards the Southern or Northern Native American components than that of USR1. Indeed, one component accounts for the majority of their ancestry; the Na Dene component for 939 and the South American components for Kennewick and Anzick1.

There is broad agreement from multiple lines of evidence, including archaeology, skeletal biology and genetics, that Native Americans entered the New World after diverging from an Asian source population [31,60,73,74]. Model-based clustering and MDS analyses show that USR1 is genetically most closely related to contemporary Native Americans, but that it is not particularly affiliated with either clade or with any major Native American group. These results suggest that USR1 diverged from the same Asian-derived source population that only later gave rise to present-day Native Americans.

9. *f3* statistics

9.1. Methods

We used outgroup *f3* statistics [75], to investigate the genetic affinity between a set of ancient Native American genomes, including USR1, and a genotype panel of contemporary worldwide populations (Section 6.1). For each ancient genome, we computed *f3* statistics of the form *f3(San; ancient, population X)* using admixtools [75]. Under the assumption that no admixture occurred in the tree *(outgroup; ancient, population X)*, drift on the lineage that leads to the outgroup (San in this case) should

remain constant regardless of the choice of the two ingroups. Thus, the value of *f3* is proportional to the degree of genetic history shared between the ancient individual and population *X*. Standard errors were estimated using the admixtools built-in weighted block jacknife procedure over 5cM blocks.

For each genome, we followed two different strategies depending on coverage depth (Section 6.2). For those ancient and modern genomes with an average depth of coverage greater than 10X, for which we were able to call genotypes with confidence (USR1, Anzick1, Athabascan1 and Aymara), we used the called genotypes for the outgroup *f3* statistics. For the remaining low depth ancient genomes with an average depth of coverage lower than 10X (Kennewick, 939 and USR2), we sampled one random allele with a base quality ≥20 and a minimum mapping quality ≥30 at the sites included in the genotype panel.

9.2. USR1 is most closely related to Native Americans but it is not part of either the NNA or SNA branches

Outgroup *f3* statistics revealed that USR1 is more closely related to present-day Native American populations than to any other population included in the panel (Figure 1a, Figure S4a). Yet, USR1 did not show increased affinity to the NNA or the SNA branch, as has been shown for other ancient Native American genomes, e.g., [44]. These exclude Na-Dene and Eskimo-Aleut-speakers, which we found to be more distant from USR1 that other Native Americans, likely due to recent Asian-derived admixture into the former (Section 10.2, Figure S5a). To determine if USR1 had a higher genetic affinity to any population, particularly relative to other Native American genomes, we compared the vector of outgroup *f3* statistics obtained for USR1 to that obtained for a set of ancient and modern Native American genomes (Figure S4b-f). In agreement with USR1 being most closely related to Native Americans, we found that USR1 and all tested Native American genomes are similarly related to Old World populations. Nevertheless, when considering Native American populations, the five tested Native American genomes yielded higher *f3* values than USR1. This pattern suggests that USR1 belonged to a population most closely related to Native Americans, but one that branched off prior to the differentiation between the two major Native American lineages. Interestingly, we observed a weak tendency (not significant, $p>0.01$ based on Z-scores from *f3* statistics) for USR1 to share less genetic drift with Eurasians than other Native Americans do (Figure S4). We explore this pattern in detail in Sections 11-13.

10. D statistics

10.1. Methods

We used *D* statistics [75] to learn about the evolutionary history of the population represented by USR1 with particular focus on the early peopling of the Americas. In

brief, the *D* statistic measures the support for a tree topology *(H1, H2; H3, H4)*, where *H1, H2, H3* and *H4* represent different populations. Under the null hypothesis *(H1, H2; H3, H4)*, the expectation is that *D* is ~0. Significant deviation from 0 may arise due to (1) the proposed tree being wrong; (2) gene flow between the lineages of the tree; or (3) differential error rates between *H1* and *H2*. We computed *D* statistics based on allele frequencies and whole genome sequence data, following the merging strategy described in Section 9.1.

Allele frequency-based *D* statistics were computed using admixtools [75], based on the genotype panel described in Section 6.1. Whole genome sequence-based *D* statistics were computed according to the 'ABBA-BABA' definition from [76], by sampling a random allele from the modern and ancient genomes described in Section 6.2. Note that even though we found USR1 and modern genomes to have similar genotyping error rates (Section 4.3), we restricted the whole genome-based analysis to transversion sites, in order to minimize potential bias introduced by aDNA damage [38]. To assess the significance of each test, we estimated standard errors for *D* using a weighted block jackknife procedure over 5cM blocks for the allele frequency-based statistics and 5Mb blocks for the whole genome sequence-based statistics. Standard errors were then used to compute *Z*-scores. *D* statistics for which $|Z|>3.3$ (which corresponds to a *p*-value of ~0.001) were regarded as statistically significant deviations from $D=0$, unless otherwise stated (Section 10.4).

10.2. Identifying Native American populations with recent Asian admixture.

The following analyses rely on a genotype panel in which Native Americans have been masked for recent European and African admixture (Section 6.1) [31]. However, it has been previously reported that some of the here-included populations bear recent Asian admixture [31,60,61,71], for which local ancestry assignment is particularly challenging due to the low genetic distance between Native Americans and East Asians. Such admixture holds the potential to bias the results, especially when exploring the relationship between Asians and Native Americans. Therefore, we defined a set of Native American populations with detectable levels of East Asian admixture. We computed *D* statistics of the form *D(Native American, Aymara; Han, YRI)* for each of the Native American populations in the panel, where the Aymara population is a generic representative of Native Americans with no Asian admixture as seen from *ADMIXTURE* results (Figure S3c). We found eight Native American populations that yielded $|Z|>3.3$ (Figure S5a). Note that in Section 10.4, we categorize Native American populations based on the Z-scores for this test and in Section 10.5, we associate significant deviations from D=0 with recent Asian gene flow into these populations.

10.3. Identifying North (NNA) and South (SNA) Native American populations.

It has been previously proposed that Native Americans can be grouped into two major clades [31,44,60], that we refer to as the North (NNA) and South (SNA) Native American

branches. We classified Native American populations and genomes (Section 6) in the reference dataset into these two clades by computing *D* statistics of the form *D(Aymara, Native American; Anzick1, YRI)*. Based on [44], we expect not to observe significant deviations from *D=0* for SNA populations. On the other hand, we expect *D<0* for NNA populations; *i.e.*, the SNA Aymara should form a clade with the Anzick1 genome to the exclusion of the test population. By using this classification criterion, we assigned 50 populations and 11 genomes to the SNA branch (Figure S5b). Together with Section 10.2, this classification serves as a basis for the interpretation of subsequent results.

10.4. USR1 and contemporary Native Americans derive from the same source population.

We examined the relationship between USR1 and the Native American, Siberian and Asian populations included in the genotype panel by computing *D* statistics of the form *D(Native American, USR1; Siberian1, Siberian2/Han)*. In what follows, we refer to Siberians and Han Chinese as 'Asians' for simplicity, unless otherwise stated. In order to minimize a potentially confounding signal arising from recent gene flow between Native Americans and northeastern Siberians, we excluded Siberian populations with documented Native American admixture (this includes Naukan, Chukchis and Koryaks) [31]. Note that we also excluded Siberian populations with less than 10 individuals, following [60]. This resulted in 18 Asian and 74 Native American populations in our sample. To account for multiple testing (11,322 possible combinations), we considered two α levels to assess the significance of the results: (1) $|Z|>3.3$, which corresponds to a *p*-value of 0.001 and (2) $|Z|>4.91$ that corresponds to a *p*-value of 0.01 after applying a Bonferroni correction for 11,322 tests, as in [60]. Since the Bonferroni correction is quite conservative, we also examined the Z-score distribution, which we expect to be approximately normal under *H0*.

In particular, we investigated two contrasting hypotheses:

H0: USR1 derived from the same source population that gave rise to contemporary Native Americans. In this case, USR1 is expected to form a clade with any Native American population to the exclusion of Asian outgroups, regardless of the choice of said outgroups. In other words, we expect *D* not to deviate significantly from 0.

H1: USR1 and Native Americans descend from different source populations; therefore, they represent different migration waves into the New World. Under this scenario, USR1 and Native Americans are expected to form clades with disjoint sets of Asian populations, which would represent the different source populations. Therefore, we would expect $D \neq 0$, and we would reject the null hypothesis of *D=0* depending on the choice of the Asian outgroups. Note that we explore the possibility of USR1 being differentially related to a particular group of Native American populations in Section 10.6.

By using the corrected significance level, we were not able to reject the null hypothesis of USR1 branching off the same source population that gave rise to Native Americans (Figures 1c and S5), except for Native American populations for which we detected Asian gene flow (Section 10.2). To better characterize the relationship between USR1 and Siberians, we assumed that the Han Chinese represent an outgroup for Siberians and Native Americans and considered *D* statistics of the form *D(Native American, USR1; Siberian, Han)*, separately. While we only rejected the null hypothesis when we considered Native American populations with recent Asian admixture (Eskimo-Aleut speakers as well as some Northern Amerind and Na Dene speakers) (Section 10.2), we found the Z-score distribution to be shifted to the right (Figure S6a). These positive Z-scores do not represent a statistically significant deviation from *D=0*, however, the systematic shift suggests that USR1 has a slight residual genetic affinity to Siberians, when compared to contemporary Native Americans. Due to its small size, this signal is not observed in more explicit models involving USR1 and other Siberian and Asian populations (Sections 17-19).

In order to confirm that USR1 and Native Americans form a clade to the exclusion of Eurasian populations (*H0*), we then computed *D(Siberian, USR1; Native American, Han)*. For all tests, we rejected the null hypothesis of USR1 forming a clade with the Siberian population with $Z>6.7$ ($p\text{-}val\sim2.08E^{-11}$), thus confirming the USR1-Native American clade (Figure S6b). These results suggest that Native Americans, especially those staying in Eastern Beringia (as represented by USR1) did not immediately become genetically isolated from the Old World, a pattern evident in the archaeology as well (Section S1). This period of gene flow has been previously documented [31,60] and we explore it in Section 18.

We then assessed whether USR1 was more closely related to specific Siberian populations by considering *D* statistics that include pairs of Siberian outgroups and not the Han Chinese. Once more we did not reject the null hypothesis of USR1 and Native Americans being equally related to Siberian populations, except for Native American populations carrying Asian admixture (Figure 1c). Moreover, the observed Z-score distribution was similar to the expected normal distribution, especially when we consider Native American populations without detectable levels of Asian admixture. Altogether, these results are consistent with USR1 and contemporary Native Americans deriving from the same source population that migrated into the Americas, a movement previously estimated as having occurred sometime after 23,000 years ago [31,77].

10.5. USR1 carries ancient North Eurasian ancestry

Based on genetic evidence, it has been shown that Native Americans derive part of their ancestry from an admixture event with an ancient North Eurasian (ANE) metapopulation, already diverged from East Asians [66]. In order to confirm that USR1 and present-day Native Americans descend from the same source population, we assessed if USR1 carried an ANE ancestry component. To do so, we used whole

genome sequence data from the 24 kya individual from the Mal'ta site in South-Central Siberia [66], and computed allele frequency-based *D* statistics of the form *D(Native American, Han; Mal'ta, YRI)*. In particular, we found that the test *D(USR1, Han; Mal'ta, YRI)* yielded a *Z*-score of -14.61, consistent with USR1 deriving a fraction of its ancestry from the ANE metapopulation (Figure S5c). Such a *Z*-score is comparable to those obtained for the rest of the Native American populations (*Z* scores ranging from -22.159 to -4.45). This result further supports the evidence that USR1 and contemporary Native Americans derived from a common source population that admixed with the ANE metapopulation in the Old World, prior to the peopling of the Americas.

10.6. USR1 is an outgroup to Northern and Southern Native Americans

Whether the Northern (NNA) and Southern (SNA) Native Americans entered the Americas as part of the same migration wave and later split, or if they represent separate migrations into the New World, is debated [31]. We used *D* statistics of the form *D(Native American, Aymara; USR1, YRI)* to determine the relationship between the USR1 genome and contemporary Native American populations (Figure 1b). To gain insight into the early divergence process of the two major Native American lineages, in light of their relationship to the USR1 genome, we tested two hypotheses:

H0: USR1 is an outgroup to both the NNA and SNA lineages. In this case, we do not expect any significant deviation from $D=0$.

H1: USR1 belonged to either the NNA or the SNA lineage. This hypothesis has two possible outcomes: (1) USR1 is part of the NNA lineage, in which case we expect statistical support for $D<0$ when the test Native American population belongs to the NNA branch; (2) USR1 is part of the SNA lineage, in which case we expect statistical support for $D>0$ when the test Native American population belongs to the NNA branch.

Using allele frequency- and whole genome sequence-based *D* statistics, we were not able to reject the null hypothesis for any of the Native American populations that were determined not to carry recent Asian admixture (Section 10.2). More specifically, we found that the NNA and SNA lineages are genetically equidistant to the USR1 genome (Figure 1b, Figure S7). These results are in line with evidence from modern genomes showing that both lineages form a monophyletic clade with respect to Old World genomes [31]. Moreover, the finding that USR1 represents a Native American population that branched off prior to the divergence of contemporary Native Americans, allows us to better characterize the NNA-SNA split. In Section 20, we assess the splits between the ancestors of USR1 and NNA+SNA, and between NNA and SNA based on these results, together with the demographic parameters estimated in Sections 18 and 19.

10.7. USR1 does not carry Australasian admixture

It has been hypothesized that the first Native Americans derive a fraction of their ancestry from a 'ghost' population related to contemporary Australasians, a hypothesis based on genetic evidence, but thought to be consistent with the so-called 'Paleoamerican' cranial morphology seen in some early American skeletal remains [78–80]. Note that all genetic studies to date have failed to link Paleoamericans to a distinctive Australasian genetic signature [31,67]. Nonetheless, it has been shown that some contemporary Native American populations have a slight genetic affinity to Australasian populations [68]. Yet, the mode and especially the chronology of the demographic processes that gave rise to this pattern remain obscure. We assessed whether USR1 is more closely related to Australasians than other contemporary Native Americans by computing *D* statistics of the form *D(Native American, USR1; Papuans, YRI)*.

In all cases, we were not able to reject the null hypothesis of USR1 and a given Native American population being equally related to Papuans (Figure S5d). This result indicates that the ancestors of USR1, did not have excess affinity to Australasians, when compared to other Native Americans. However, with the data at hand we cannot rule out that the reported signal originated from an admixture event that took place during the initial peopling of the Americas. While USR1 represents a relict population of the Native American ancestors, strong genetic structure in the ancestral population might have caused a differential genetic affinity to Australasians among Native American populations. Ultimately, additional ancient genomic data, particularly from South America where the signal is most noticeable among modern groups (notably the Surui), could prove useful to resolving this issue.

11. Enhanced D statistics

When comparing vectors of outgroup *f3* statistics obtained for USR1 and other Native American genomes (Section 9), we found these vectors to be highly correlated for Old World populations (Figure S4b-f), consistent with a shared source population for USR1 and Native Americans. However, we observed a weak tendency (not significant, $p>0.01$ based on Z-scores from *f3* statistics) for USR1 to share less genetic drift with Eurasians than seen in other Native Americans (Figure S4). In order to investigate whether USR1 or Native Americans carried a distinctive admixture component, thus explaining the offset in pairwise *f3* statistics, we computed 'enhanced' *D* statistics [64] based on different ascertainment schemes.

11.1. Methods

'Enhanced' D statistics of the form $D_{enhanced}(H1, H2; H3, Outgroup)$ differ from 'basic' *D* statistics (Section 10) in that the analysis is restricted to sites in which a set of 'outgroup' genomes carry the same allele, and *H3* is homozygous for a different allele [64]. This ascertainment step enriches for variants that arose in the lineage leading to *H3* and excludes ancestral variation present before the split between *H1* and *H2*, thus

'enhancing' any potential signal produced by gene flow. Note that we use the term 'outgroup' genomes to represent populations that are symmetrical (an outgroup) to the clade formed by H1 and H2; said genomes do not necessarily form a clade to the exclusion of H3. Rather, we group them under the 'outgroup' term to indicate that they carry an allele different from that in H3. After selecting such sites for each comparison, we computed *D* statistics and assessed their statistical significance as described in Section 10 for whole genome sequence-based tests. Although the USR1 genotypes are based on high depth sequencing of USER treated libraries [43](Section 2), we also filtered out transition polymorphisms in order to reduce the bias introduced by substitutions characteristic to aDNA [38]. We note that results from both approaches are highly concordant, further supporting the observation that the genotyping error rate in USR1 is comparable to that of modern genomes (Section 4.3).

11.2. Testing for Australasian ancestry in USR1

In Section 10.7 we concluded that USR1 did not carry additional Australasian ancestry, compared to other Native Americans. However, such ancestry has been found in low amounts in certain Native American groups [31,68]. We computed enhanced *D* statistics of the form *D(Native American, USR1; Papuan, Outgroup)* to assess if USR1 carried Australasian admixture not detectable with 'standard' *D* statistics. To do so, we defined three incremental sets of genomes that represent outgroups to the *(Native American, USR1)* clade:

A: 5 African genomes (Yoruba, San, Mandenka, Dinka and Mbuti)
B: 5 African + 2 European genomes (French and Sardinian)
C: 5 African + 2 European + 2 Asian genomes (Han and Dai)

In what follows, we refer to this test as the 'Australasian test'. Due to a more efficient enrichment for Papuan-specific variants, increasing the number of 'outgroup' genomes resulted in higher values of D. When we considered outgroups A and B, we were not able to reject the null hypothesis of USR1 and Native Americans being symmetrical to the Papuan genome. However, when we considered outgroup C, we found statistical support for $D>0$ ($|Z|>3.3$) for some tests (Figure S8a). Although the latter was not the case for all *H1* genomes, likely due to the low number of remaining 'ABBA-BABA' sites after the ascertainment step using outgroup C, we observed $D>0$ in most cases (but not statistically significant under the $|Z|>3.3$ criterion). The Aleutian genome, which has been shown to carry an Asian genetic signature [31] was the only one to yield negative values of D. While these results suggest that the USR1 genome is slightly closer to the Papuan individual than are other Native Americans, they depend on a strict ascertainment scheme. Moreover, pairwise *f3* statistics showed that USR1 does not have a higher genetic affinity for Papuans than other Native Americans (Figure S4). In fact, USR1 yielded lower *f3* values for all non-African populations (including Papuans), than other Native Americans. Thus, in what follows we explore a different source of admixture into USR1.

11.3. Testing for archaic hominin ancestry in USR1

We then tested whether USR1 and other Native Americans carried different proportions of archaic hominin admixture. In principle, a larger archaic ancestry proportion could cause USR1 to systematically appear more distant to all Old World modern populations; thus explaining the pairwise *f3* statistics offset. We tested this scenario by computing enhanced *D* statistics of the form *D(Non-African, USR1; Archaic, Chimp)* on sites where the five African genomes and the chimp carry the same allele, using the high coverage Altai Neanderthal [65] and Denisovan [64] genomes. In what follows, we refer to these tests as the 'Neanderthal' and 'Denisovan' tests, respectively.

We were not able to reject the null hypothesis of USR1 and other non-Africans being equally related to the Altai Neanderthal genome (Figure S8b), except for the Papuan individual, which has been shown to carry larger archaic admixture [65]. On the other hand, when we tested for Denisovan ancestry, we observed a trend for *D* to be positive (but not statistically significant in most cases) when considering Native Americans for both 'basic' and 'enhanced' *D* statistics (Figure S8c). This pattern suggests that USR1 carries more Denisovan ancestry than seen in other Native Americans. Genomic evidence has shown that anatomically modern humans diverged from Denisovans before migrating and diversifying out of Africa [64]. Therefore, excess Denisovan ancestry could be put forward as a potential cause for the observed pairwise *f3* statistics patterns in which USR1 shares less drift with Eurasians than other Native Americans do. Note that the Huichol genome yielded *D(Huichol, USR1; Denisova, Chimp + 5 Africans)= 0.003, Z= 0.148*; further inspection of outgroup *f3* statistics of the form *f3(San; Huichol, X)* revealed a pattern similar to that of USR1, when compared to the Aymara genome (Figure S13k). In what follows, we characterize the Denisovan ancestry component in USR1 and conclude that this pattern is most likely not due to an independent admixture event into the USR1 lineage but a consequence of 'heterogeneous' Denisovan admixture proportions in Native Americans.

11.4. Australasian ancestry as a function of Denisovan ancestry in USR1

It has been shown that Papuans are among the modern human populations that carry the largest fraction of Denisovan ancestry [64]. Therefore, we investigated if Denisovan admixture could drive the apparent excess Australasian ancestry in USR1. Under such scenario, we expect the 'Australasian' and 'Denisovan' tests to have the same sign and to be similar in magnitude. We computed both enhanced *D* statistics by using two sets of genomes that represent outgroups to the *(Native American, USR1)* clade:

A: 5 African (Yoruba, San, Mandenka, Dinka and Mbuti) + 2 European (French and Sardinian) + 2 Asian genomes (Han and Dai)
B: Chimp + 5 African + 2 European + 2 Asian genomes

We found USR1 to yield similar positive *D* values for both tests, regardless of which Native American genome was tested. We compared these results to those obtained for other Native American genomes (Anzick1, Aymara and Surui) (Section 6). In contrast to USR1, the Anzick1 and Aymara genomes produced non-significant positive and negative results for both tests. In addition, the Surui genome consistently yielded positive values for the 'Papuan test', independently of the sign of the 'Denisovan test' (Figure S9a). These results suggest that the higher genetic affinity between certain Amazonian groups and Australasians [31,68] is independent of archaic hominin ancestry. On the other hand, the apparent Australasian ancestry in USR1 could be explained by excess Denisovan ancestry, rather than by gene flow between the ancestors of USR1 and an ancient population related to present-day Australasians.

11.5. The Denisovan genome is the best proxy for the unknown ancestry in USR1

To further confirm if the Denisovan genome is the best available proxy for the 'unknown' ancestry in USR1, we studied *D* statistics of the form *D(Aymara, USR1; H3, Han)*, where *H3* represents the French, Papuan, Yoruba, Altai Neandertal, Denisovan and chimp genomes. We assumed that the Aymara and USR1 genomes form a clade and that USR1 carries an ancestry signal from an 'introgressing population' related to *H3* (compared to Aymara). In this case, *D* is proportional to the length of the branch between the common ancestor of the 'introgressing population' and *H3* on the one hand, and the common ancestor of Han, Aymara and USR1 on the other. It is on this branch that the genetic drift paths *H3-Han* and *Aymara-USR1* overlap (via the admixture edge) [75], thus contributing to *D* (see Figure S10 for a schematic representation of this test). Consequently, we expect *D* to increase with the genetic relatedness between *H3* and the true 'introgressing population'. Note that for this test we assume that admixture events into the branch leading to Han, Aymara and USR1 affect all three equally, *e.g.* Neanderthal admixture into the out of Africa population [64,65]. Thus *D* will deviate from 0 when USR1 and Aymara are not symmetric to *H3*, which could be a consequence of both genomes carrying different admixture proportions from an admixture event before their split, or an independent admixture event into either genome.

In agreement with the enhanced *D* statistic results, the Denisovan genome yielded the highest value for *D(Aymara, USR1; H3, Han)*, followed by the Altai Neanderthal (Figure S9b). Note that this value is larger for the Denisovan genome than that obtained for the chimp, which is the most divergent genome in the set, thus suggesting that this signal is not driven by sequencing or genotyping errors (see Section 4.3 for genotyping error estimates). We note that we did not observe any statistically significant deviation from *D=0*, likely due to the Denisovan component in USR1 being small, thus hard to detect with single-genome-based *D* statistics. However increasing values of D are consistent with the overlap in drift paths shown in Figure S10, which suggests that the distinctive ancestry component carried by USR1 is most closely related to the Denisovan genome.

11.6. Similar differences in Denisovan ancestry are observed in random pairs of Papuan genomes

When testing for Denisovan ancestry in USR1 (Section 11.3), we observed positive values of the enhanced test *D(Native American, USR1; Denisova, Chimp)* for all Native American genomes. However, we did not detect statistically significant deviations from *D=0*, based on a $|Z|>3.3$ criterion, for most comparisons (Figure S8c). Given the small size of the signal, we investigated whether this pattern could arise by randomly sampling an individual carrying a larger Denisovan ancestry component from a population with already variable Denisovan admixture proportions. To do so, we computed *D(Papuan Papuan; Denisova, Chimp)* on sites where the five African genomes and the chimp carry the same allele, for all 10 possible pairs of Papuan genomes (Table S3).

For all tests, we were not able to reject the null hypothesis of both Papuan individuals being equally related to the Denisovan genome (Figure S9c). However, we obtained values of D as extreme as -0.049, corresponding to a Z score of -2.417. These values are comparable to those obtained for Native American-USR1 enhanced *D*-statistics, *e.g. D(Aymara, USR1; Denisovan, Chimp)*=0.0486, *Z*=2.12. These findings suggest that the excess Denisovan ancestry in USR1 is comparable to that of a randomly sampled individual from a population with variable Denisovan admixture proportions.

12. Denisovan ancestry tracts in the USR1 genome

12.1. Methods

By using *D* statisitics, we found a slight excess Denisovan ancestry in USR1, compared to other Native Americans such as the Aymara (Section 11). To determine if such ancestry derives from an admixture event independent of the already characterized introgression(s) into anatomically modern humans [64], we inferred Denisovan ancestry tracts in USR1 and the reference genomes (Table S3), and examined the tract length distributions qualitatively. If the ancestors of USR1 obtained the Denisovan-related genetic component through a more recent admixture event that postdates the split from Native Americans, we expect to find longer Denisovan ancestry tracts in USR1 [81].

We followed a strategy similar to that used in [82]. We first defined a set of sites where the high coverage Denisovan genome [64] carries a derived allele and all the African samples in the 1000 Genomes project [83] carry the ancestral allele. In order to minimize potential noise arising from shared ancestry between Neanderthals and Denisovans, we required the high-coverage Altai Neanderthal genome [65] to be homozygous for the ancestral allele at such sites. Finally, we defined Denisovan-derived tracts as stretches containing at least five consecutive Denisovan alleles and allowed for a maximum gap size between Denisovan alleles of 100 kb.

12.2. Denisovan ancestry tracts in USR1 do not support additional archaic gene flow into USR1

We compared the Denisovan ancestry tract length distribution in USR1 to those obtained for a set of reference genomes (Figure S11). By summing the lengths of the inferred tracts, we found USR1 carries ~3.6 Mb of Denisovan-derived DNA, while other Native Americans carry between 1.4 and 3.4 Mb. Additionally, we found comparable mean tract lengths for USR1 (~0.082 Mb) and other Native American genomes (0.049-0.093 Mb). These results support the idea that USR1 carries more Denisovan-derived ancestry than other Native Americans present in the whole genome dataset (Section 11). Moreover, they suggest that the Denisovan admixture component in USR1 is not a product of a separate admixture event, since it is not distributed in longer tracts than those inferred for other Native Americans. Note that under the scenario of an independent admixture event into the USR1 lineage after divergence from Native Americans, the expectation is to find longer Denisovan ancestry tracts in USR1, as this would allow a shorter time for recombination to break down contiguous introgressed Denisovan segments [81].

We then considered the Denisovan ancestry tract length distributions in five modern Papuan genomes. In agreement with previous results [64], we found Papuans to have the largest Denisovan ancestry proportions (13.5-16.8Mb) and mean tract length (0.098-0.106Mb), among the tested genomes. Interestingly, the Denisovan ancestry proportion range in Papuans is also larger than that observed in Native Americans. Together with Section 11.6, these results support that the signal observed in USR1 could be caused by sampling an ancient genome from a population with heterogeneous Denisovan ancestry proportions.

13. Identifying USR1-specific tracts in the USR1 genome

13.1. Methods

We took a second approach for investigating the pairwise *f3* statistics offset resulting from comparing USR1 to other Native American genomes (Sections 9 and 11, Figure S4). We identified segments of the USR1 genome that are unique to it with respect to other Native Americans. We hypothesize that these segments contain the ancestry responsible for the observed outgroup *f3* statistics patterns. We followed the same strategy used for Denisovan ancestry tracts inference (Section 12) [82]. In this case, inference was based on '*X*-specific' sites, defined as sites where a genome *X* carries an allele not present in any of the other 18 Native American genomes in the reference dataset (Table S3). We did not include the Aymara genome in the reference set, since we also inferred Aymara-specific tracts in the Aymara genome, as a control.

13.2. Assessment of the USR1-specific tracts

We found 116,936 USR1-specific sites and 56,520 Aymara-specific sites that we then used for inferring '*X*-specific' tracts for each genome. '*X*-specific' sites may arise due to: (1) admixture from a non-Native American source; (2) lineage-specific mutations; (3) genotyping errors; or (4) 'misidentification' due to underrepresentation in the limited reference genome set. Due to recombination, we expect admixture-derived '*X*-specific' sites to cluster together in the genome [81]. Therefore, we analyzed '*X*-specific' tracts based on their '*X*-specific' site density. For each tract, we defined its density as the number of '*X*-specific' sites in the tract divided by the tract length. We then defined 18 tract sets based on the deciles (10%-90%) of the density value distribution. For each decile, we defined two sets of tracts; one containing 'dense' tracts, with density greater than the decile, and a second set containing 'sparse' tracts, with density lower than the decile (Figure S12a).

We then compared the tract length distributions obtained for the USR1 and Aymara genomes (Figure S12b). For all density cutoff values, we found USR1 to carry more specific tracts that are on average longer than those in the Aymara genome. Note however, that in all cases we obtained an Aymara-specific tract length distribution similar to that of USR1. These results suggest that by using this method, we not only identify admixture-derived '*X*-specific' tracts, but also 'noise'-derived tracts that most likely include lineage-specific mutations, genotyping errors, or spurious '*X*-specific' sites.

13.3. Denser USR1-specific tracts contain Denisovan-derived ancestry and are responsible for the pairwise *f3* statistics offset

We masked the USR1 genome according to the different density-based tract sets and computed *f3(San; USR1, X)*, as detailed in Section 9 and compared the *f3* vectors to that obtained for *f3(San; Aymara, X)*. We first masked the 'sparse' tract sets (tracts with density lower than a given decile) and observed little to no effect when we did not mask the densest tracts, *i.e.*, tracts with density greater than the 70% decile (Figure S13a-e). Accordingly, when we masked the 'dense' tract sets, we found that masking the 50% densest tracts already reduced the observed offset notably (Figure S13f-j). Based on these results, we took a conservative approach and only considered the tracts denser than 50% decile (90.171 Mb and 7,037 sites of the SNP array data) for the following test.

We used three different approaches to confirm that USR1-specific tracts contain excess Denisovan ancestry. First, we computed enhanced *D* statistics of the form *D(Aymara, USR1; Denisova, Chimp+5Africans)* (Section 11). We separately considered USR1 masked on the 50% densest tracts, and the 50% densest tracts. These tests revealed that masking USR1-specific tracts effectively cancels the Denisovan signal in USR1 (*D=-0.0013, Z=-0.057*). Moreover, we observed a highly significant deviation from *D=0* when we considered the USR1-specific tracts separately, even though this test was based on ~500 ABBA+BABA sites (*D=0.544, Z=9.5*). We then computed *f3(San; X,*

USR1) by randomly masking 7,037 sites in USR1 and observed no effect in the offset, when compared to *f3(San; X, Aymara)* (Figure S13l). Finally, we compared the number of Denisovan-derived alleles (as defined in Section 12) carried by the USR1-specific tracts and a set of randomly sampled tracts with the same length distribution. USR1-specific tracts overlapped with 1,890 Denisovan ancestry informative markers and carried 205 Denisovan-derived alleles. On the other hand, the randomly sampled tracts overlapped with 2064 informative sites and carried 42 Denisovan-derived alleles. Based on these counts, we performed a χ^2 test of homogeneity for which we rejected the null hypothesis of USR1-specific and randomly sampled tracts having similar proportions of Denisovan alleles, with $p < 2.2e^{-16}$.

Together with results shown in Sections 11 and 12, these results confirm that the USR1 genome bears excess genetic ancestry from a population related to the Denisovan genome, relative to other Native Americans. This ancestry is in turn responsible for the pairwise *f3* statistics offset observed when comparing USR1 to other Native Americans. Moreover, the excess Denisovan ancestry in USR1 is not consistent with an independent admixture event, but rather resulted from the random sampling of an individual from a population with heterogeneous Denisovan ancestry proportions.

13.4. Excess Denisovan ancestry in USR1 explains the pairwise f3 statistics offset: a simulation study.

We carried out a simulation study to confirm if excess Denisovan ancestry in USR1 could cause the pairwise f3 statistics offset observed in Figure S4. We simulated data under the demographic model shown in Figure S14a using *msprime* [84], assuming a mutation rate of $\mu = 2.5\times10^{-8}$ bp/gen, a recombination rate of $\rho = 1\times10^{-8}$ bp/gen, and a generation time of 29 years. The model includes five populations that correspond to Denisovans, Africans, Aymara, the population represented by USR1 and a population X that represents Eurasian populations. Demographic parameters were based on the parameters inferred in Sections 18 and 19 and [64,65,85]. We set population split times to 450 kya for the Denisovan split (T_D), 72 kya for the African split (T_A) and 20 kya for the split between Aymara and USR1 (T_{Ay}). We assumed constant effective population sizes of N_D={100, 400, 700, 1000} for Denisovans, N_A=20,000 for Africans, N_B=1,600 for the out of Africa population, N_{Ay}=1,500 for Aymara, and N_U=2,000 for USR1. For population X we considered a series of different split times T_X={23,000, 24,500, ..., 50,000}, and three different population sizes N_X={3000, 8000, 13000}, meant to represent different Eurasian populations. Finally, we considered two admixture events from Denisovans: one into the out of Africa population 45 kya (T_{adm}) and a second one into USR1 16kya (T_{adm2}). For the former we considered an admixture proportion of p_{adm}=0.25% and for the latter we simulated data with different proportions p_{adm2}={0, 0.0033, 0.0066, 0.01, 0.015}. Note that west Eurasian populations that diverged prior to the inferred Denisovan admixture and east Eurasians [64] are taken into account in the cases in which $T_X > T_{adm}$.

For each combination of N_D, p_{adm2}, T_X, N_X, we simulated a total of 22 diploid individuals: ten diploid individuals from the African and X populations, one diploid individual from the Aymara population and one diploid individual from the USR1 population sampled 11.5 kya. Each individual contained 20 independent 100 Mb chromosomal segments, that we thinned to a segregating site density similar to that of the SNP array dataset (one SNP per ~14.5 kb, Section 6.1). We then computed *f3(African; X, USR1)* and *f3(African; X, Aymara)*, and compared both statistics as detailed in Section 9. We summarize these results in Figure S14b.

As expected, *f3* was proportional to T_X and the choice of N_X had no major effect on the statistics [75]. In addition, we observed little effect of N_D on the estimates. For all simulated datasets that included an admixture event into the USR1 lineage, we observed a pairwise *f3* offset in the same direction as that observed for the SNP array data (Figures S4 and S14b). Yet, models with the highest values of p_{adm2} yielded the results that were the most comparable to those obtained for the array data. In contrast, we did not observe a similar pattern for simulations without such admixture. These results show that an additional Denisovan-related ancestry component in USR1 could give rise to the observed offset in pairwise *f3* statistics comparisons.

We note that the observed variance for each point in Figure S14b is a consequence of the simulation setup. For each point, we simulated the demographic scenario *((((Aymara, USR1), X), African), Denisovan)*, independently. Thus, for each different instance of *X* (combination of N_X and T_X), different Aymara and USR1 genomes were simulated. Provided the small size of p_{adm2}, and the short time comprised between T_{adm2} and the USR1 sampling time, we expect some degree of variance in the 'Denisovan signal' in USR1 for each data point. To illustrate this, we carried out a similar simulation experiment, but in this case we jointly simulated the 'African' population, the single 'USR1' and 'Aymara' individuals, and 51 instances of *X*. The latter were assigned random effective population sizes (N_{Xi}) between 3,000 and 13,000 and split times (in years) $T_{Xi}=\{25{,}000, 25{,}500, ..., 50{,}000\}$. For this new setup in which only one 'USR1' and one 'Aymara' genome were simulated for all instances of *X*, we observed less variance on the f3 statistics (Figure S15a), and the results are qualitatively comparable to those obtained for real data (Figure S4).

Based on three independent replicates of this experiment, we observed that although the 'offset' is similar for all instances of *X*, the size of the 'offset' is not well predicted by p_{adm2}, unless the three replicates are pooled together (Figure S15b). In contrast, we observed a better correspondence between the simulations presented in Figure S14b and p_{adm2}. We interpret that in the latter case, multiple simulated instances of USR1 allow for the variance in the 'Denisovan signal' to be averaged out, and thus we capture the effect of p_{adm2} more effectively. While we consider the approach in Figure S14b to be robust enough for illustrating how a slight excess of Denisovan admixture in USR1 can produce the observed signal, we caution that random sampling may produce a similar signal. Yet, these results support the idea that slightly increased Denisovan ancestry in

USR1 will, in average, produce a signal similar to that observed in the real data. Based on our analysis of Denisovan ancestry tract length distributions (Section 12), we reiterate that increased Denisovan ancestry in USR1 is most likely a consequence of randomly sampling an ancient genome from a population with heterogeneous small Denisovan ancestry proportions, and not of an independent admixture event into USR1.

We emphasize that while we simulated additional Denisovan ancestry in USR1 as an independent admixture event, a comparison of the length distribution of Denisovan ancestry tracts in USR1 and other Native Americans indicate that such an event did not take place between the time in which USR1 split from other Native Americans and 11.5 kya (Section 12). Rather, the excess Denisovan-derived ancestry in USR1 most likely resulted from sampling an ancient genome from a population with heterogeneous Denisovan ancestry proportions (Section 11), likely affected by natural selection and differential drift [86]. Therefore, our simulations are only meant to show the level of differential admixture that could explain the results.

14. Treemix

14.1. Methods

We then explored the phylogenetic placement of USR1 in the broader context of Old and New World populations. To do so, we inferred maximum likelihood admixture graphs using *TreeMix* [87]. For this analysis, we considered a subset of the populations represented by modern and ancient high-coverage genomes (Section 6), which includes all non-Africans with a sequencing depth greater than 10X, the high coverage Denisovan and African San genomes, as well as the 24,000 year old ~1X Mal'ta genome. We only considered transversion sites to minimize the bias introduced by aDNA damage in the Mal'ta genome, for which one read was sampled at every variable site, which resulted in 631,192 sites. We ran *TreeMix* assuming zero to five migration edges, allowed a final rearrangement of the tree after the last fitting step, and disabled the sample size correction procedure as a number of the populations in both dataset consist of a single individual. To account for linkage disequilibrium in the SNP-dense dataset, we grouped the sites in blocks of 1,020 SNPs, which corresponds to ~5 Mb blocks. For each number of migrations, we ran 1,000 replicates with random seeds and kept the run with the highest likelihood.

14.2. USR1 is basal to North and South Native Americans

When fitting a tree with no migration edges, USR1 was placed in a basal position with respect to Native Americans, including Athabascan-speakers (NNA) as well as Native Americans from Mexico and South America (SNA) (Figure S16). While USR1 remained an outgroup for Native Americans, serial addition of migration edges showed admixture events between: (1) Ancient North Eurasians (as represented by the Mal'ta genome) and the Native American ancestral population (including USR1); (2)

Denisovans and Papuans [64]; (3) Papuans and East Asians; (4) Greenlandic Inuits and Siberian Koryaks and Saqqaq; and, finally, (5) a basal human population and East Asians.

These results are in agreement with those obtained using *f3* (Section 9) and *D* statistics (Section 10), in that USR1 is a Native American population that is not more closely related to either NNA or SNA. Contemporary Native Americans have been shown to derive approximately one third of their ancestry from an ancient North Eurasian metapopulation [66]. Together with the results in Section 10.5, this analysis shows that the USR1 genome already carries Mal'ta related ancestry, suggesting that the admixture event that gave rise to Native Americans occurred prior to the USR1-Native American split. Moreover, the residuals after fitting 5 admixture edges do not provide any strong indication for additional migration between the population represented by USR1 and any other modern or ancient population included in this analysis. This observation further supports that the particularly increased Denisovan ancestry in USR1 is not due to an additional admixture event into the USR1 lineage (Sections 11-13).

15. Pairwise branch lengths

15.1. Methods

We then performed pairwise comparisons between a set of high coverage genomes including USR1 (Section 6) and measured the amount of drift in each branch after the split. To do so, we used the method described in [44], which was originally used as a test for direct ancestry, and which corresponds to the case of no drift in one of the branches. In brief, we consider the split between two diploid genomes with genotypes *{0,1,2}* and five possible genotype configurations *{00, 01, 10, 11, 02}*. The probability of each configuration can be expressed in terms of five parameters; namely, the probability of coalescence in any of the branches before the divergence and three parameters representing the folded site frequency spectrum of the four chromosomes in the ancestral population. Based on the observed counts for each genotype configuration, we obtain maximum likelihood estimates for the five parameters of the model, through numeric optimization. Note that we restricted this analysis to sites that are variable in five African genomes (Section 6) to enrich for mutations that arose prior to the split between both genomes [88]. Additionally, given that the Anzick1 genome has a slightly higher 'genotype error rate' (Section 4.3), we only considered sites with a minimum depth of coverage of 15X.

15.2. USR1 and Native Americans are equally drifted from the Han genome

In Section 10, we concluded that the USR1 lineage derived from the same source population that branched off Asians sometime after 25,000 years ago and that gave rise to present-day Native Americans [31,77]. Under this scenario, the expectation is that the

drift on the branch leading to East Asians from the East Asian-Native American split will be similar to that of East Asians after the East-Asian-USR1 divergence. To test this, we estimated the amount of drift in the Han Chinese branch (representing East Asians), after the split from a set of high-coverage genomes including USR1 (Section 6).

In agreement with the known genetic relationships of East Asians, we observed a larger amount of drift in the Han branch after divergence from Europeans, followed by Papuans, Native Americans, Siberians and the Dai Chinese [89] (Figure S17a). In particular, all comparisons with Native Americans including USR1, yielded similar estimates for the length of the branch leading to the Han genome. These estimates provide further support for USR1 and present-day Native Americans having been part of the same source population that diverged from Asians before entering the New World.

15.3. North and South Native Americans are equally drifted from the USR1 genome

We followed a rationale similar to the above in order to learn about the relationship between USR1 and other Native Americans. In previous sections, we inferred that North and South Native Americans are equally distant from USR1. Under this hypothesis, we expect the branch leading to USR1 to have a similar amount of drift when compared to other Native Americans. We obtained maximum likelihood estimates (MLEs) for the length of the branch leading to USR1 relative to the split from a set of high-coverage genomes (Section 6). As expected from a Native American genome, we found the branch leading to USR1 to be longer after divergence from Europeans, followed by Papuans, East Asians, Siberians and the rest of the Native American genomes [89] (Figure S17b). Moreover, the branches leading to USR1 after divergence from North (Athabascans) and South Native Americans yielded drift estimates that were similar in value to each other. This result is consistent with USR1 being equally related to both NNA and SNA, thus confirming the placement of USR1 as an outgroup to both lineages. Note that the Athabascan genomes that were analyzed bear ~10% Asian-derived admixture; therefore, we warn that this method has low resolution for differentiating restricted admixture between genetically close populations such as Asians and Native Americans. We explore the source of this Eurasian admixture into Athabascans in Section 17.

16. Genomic divergence

16.1. Methods

To confirm the results obtained through *D* statistics, we estimated the average DNA divergence between pairs of genomes *(A,B)*, as was done in [90]. In brief, we examined tree topologies of the form *(O; A ,B)* and recorded the number of differences (*nA, nB, nO*) that are unique to each lineage. Then, we estimated the DNA divergence between *A*

and *B*, relative to *O* to be *Div(O; A,B)=nA/((nA+nO)/2)*. More specifically, by assuming constant evolutionary rates along the three lineages, *Div(O; A,B)* corresponds to the ratio of the length of the branch that leads to *A* from the common ancestor of *A* and *B,* and the length of the branch that leads to *A* from the common ancestor of *A*, *B* and O. Note that these estimates correspond to average DNA divergence, which necessarily occurred before the actual population split. Moreover, a direct translation between relative divergence estimates and an actual estimate for the time to divergence is challenging due to the uncertainty associated to the dating of the split between African and non-African populations, *e.g.*, [91,92]. Therefore, we only considered the relative estimates qualitatively, to learn about the evolutionary history of USR1. We set *O* to be the genome of a Yoruba individual and estimated *Div(O; A,B)* for pairs of genomes *(A,B)* by sampling one random allele from each of the genomes at sites where all three genomes had non-missing genotypes. Note that we restricted the analysis to transversion polymorphisms, in order to reduce potential bias from aDNA damage. We then built 95% confidence intervals for the point estimates based on standard errors obtained through a weighted block-jacknife procedure over 5Mb blocks, similar to the one implemented for *D* statistics.

16.2. USR1 and Native Americans have similar average DNA divergence from East Asians.

By using *D* statistics, we found that USR1 and contemporary Native Americans derived from the same source population that branched off East Asians around the time of the Last Glacial Maximum [31]. Under this model, we expect the East Asians divergence from USR1 to be similar to that from SNA and NNA. To test this hypothesis, we computed the average DNA divergence between a Han Chinese individual (representing East Asians) and a set of genomes including two Athabascan genomes representing the NNA branch, as well as 13 genomes belonging to the SNA branch (Section 6).

In agreement with the dispersal of anatomically modern humans out of Africa [89], we found that the Han divergence from Native Americans (including USR1) was greater than that from the East Asian Dai genome (69.45%), but less than that from two European genomes (80.3% in average). In particular, we found that the Han-USR1 (73.1%), Han-SNA (72.13-72.92%) and Han-NNA (72.27- 72.43%) divergence estimates produced overlapping confidence intervals (Figure S18a). These results support the affiliation of USR1 to the same source population from which NNA and SNA derived.

16.3. USR1 is equally diverged from South and North Native Americans.

Through *D* statistics, we also showed that both NNA and SNA are equally related to USR1. Considering the age and geographic location of the USR site, this result implies that USR1 represents a relict of the population from which the NNA and SNA lineages both diverged. Under this scenario, the expectation is that both the NNA and SNA

lineages should have a similar genetic divergence from USR1. To test this hypothesis, we computed the average DNA between a set of modern genomes (Section 6) and the USR1 genome.

We found that Native American genomes are the least diverged from USR1, followed by Asians, Europeans and Papuans. More specifically we observed SNA-USR1 (66.6-67.65%) and NNA-USR1 (67.42-67.98%) divergence estimates have overlapping confidence intervals (FigureS17b). Together with the archaeological context of the USR site and results shown in Section 10.5, this result provides further support for a common Asian source for the NNA and SNA branches, and that both branches most likely split in North America, South of Beringia. Similar to Section 15, we note that this method has little resolution for differentiating the Asian-derived admixture signal into Athabascans (10%). We explore the source of this Old World admixture in Section 17

17. Assessing the origin of Na-Dene and Inuit

Based on linguistic and anthropological evidence, it has been suggested that some Native American groups, namely Athabascans (Na-Dene-speakers) and Inuit, derive at least part of their ancestry from secondary migrations from the Eurasia [8,63,93]. These claims have found support from modern and ancient genetic studies [61,71]. However, the source of these migrations is still debated. It has been hypothesized that Athabascans share a linguistic link with the Yeniseian-speaking Kets from Siberia [94,95]. Moreover, genetic studies have suggested that this link was mediated by the Palaeo-Eskimo expansion via the American High Arctic ~6 kya [96,97]. Importantly, this expansion has been proposed to be distinct from the migration that gave rise to contemporary Inuit [60,71]. Additionally, a genetic study put forward that Athabascans derive their Native American ancestry from the NNA branch, while Inuit are most closely related to the SNA branch [60].

17.1. *Treemix* approach

We approached this question by fitting admixture graphs using a dataset containing modern and ancient whole genomes. Besides the 'admixed' Native American groups (Greenlandic Inuit and Athabascans), these include Siberians, East Asians and Papuans, an African outgroup (Yoruba), as well as the Denisova, Anzick1, Saqqaq, Mal'ta and USR1 ancient genomes (Section 6). First, we used the heuristic approach in *Treemix* [87] to identify the Siberian or Asian populations that potentially contributed to the admixed Native Americans. We obtained maximum likelihood (ML) graphs by following the strategy described in Section 14, and added migration edges until both, Greenlandic Inuit and Athabascans, were admixed. Additionally, we estimated the support for internal nodes and migration edges through a bootstrap procedure. For each dataset, for each number of migrations, we generated 200 independent pseudo-replicates by randomly sampling 5Mb blocks, with replacement. For each replicate, we ran one hundred optimizations with random seeds in order to obtain ML solutions.

We observed migration edges from the Saqqaq individual and Koryaks into Inuit and Athabascans, respectively (Figure S19). However, the placement of East Asians in the tree was complicated, likely due to overrepresentation of Siberian ingroups bearing different proportions of ancient north Eurasian ancestry, *e.g.*, the Nivkhs. Therefore, we repeated this analysis on a reduced dataset from which Buryats, Yakuts and Nivkhs were excluded, as they were not found to contribute ancestry to the admixed Native Americans in previous runs. Note that we did not exclude Kets from the dataset, in order to explore the proposed Dene-Yeniseian link [94,98]. While accurately placing East Asians, the results obtained with the reduced dataset confirmed that the Saqqaq individual and the Koryaks were the best proxies for the Asian components in admixed Native Americans (Figure S20).

17.2. *qpGraph* approach

We confirmed these results by fitting admixture graphs based on *f*-statistics [60,75]. We compiled a dataset containing the following genomes: Yoruba (*n=1*), Mal'ta (*n=1*), Han (*n=1*), Ket (*n=2*), Koryak (*n=2*), Saqqaq (*n=1*), USR1 (*n=1*), Anzick1 (*n=1*), Aymara (*n=1*), Athabascan (*n=2*) and Greenlandic Inuit (*n=2*). We first produced a 'seed' graph modeling the Yoruba, Han and Mal'ta genomes as non-admixed outgroups, and the Kets and non-admixed Native Americans as deriving from two independent admixture events between East Asians (EA) and ancient north Eurasians (ANE) (*maximum |Z|=0.42*). We then added Koryaks, Saqqaq, Athabascans and Greenlanders, sequentially, as admixed leaves. For each admixed leaf, we enumerated all possible edge pairs using the tools from [99], tested them using *qpGraph* and kept the pair that produced the graph with the best maximum |*Z*| and fitting scores. We tested the robustness of this method by adding the admixed groups in a different order, and arrived to a similar result in all cases.

In agreement with *treemix* results, the best graph (*maximum Z=3.269; p-val > 0.001*) showed that Koryaks, Eskimos, Native Americans and Kets could be modeled as deriving from different mixtures between EA and ANE (Figure 3). Interestingly the ANE component in Koryaks and Eskimos appear to be very similar (0 and 1-drift edges) while the ANE ancestry in Native Americans and Kets is more diverged. Contrastingly, the EA populations giving rise to these groups are more divergent. This result confirmed the *treemix* result that Athabascans derive from an admixture between NNA and an Asian source, most closely related to the Asian component in present-day Koryaks. On the other hand, Inuit derive from an admixture event between NNA and a Siberian population most closely related to Palaeo-Eskimos. Interestingly, short branches separate the Asian populations giving rise to Koryaks and Athabascans, in contrast to the Saqqaq and the Siberian source for Inuit. While this pattern may be a consequence of a higher genotype error rate in the Saqqaq genome, it could also be a reflection of Palaeo-Eskimos and Neo-Eskimos being the end-product of two different migrations into the Americas after a deeper split in Siberia [60,71].

17.3. Robustness of the model

Since our model is based on ancient and modern genomes each with different sequencing and genotyping error profiles, we assessed its robustness by computing *D*-statistics using an extended dataset. We included the UstIshim genome [69] that served as an aDNA-based non-African outgroup, as well as different versions of the Saqqaq genome, which allowed us to control for its unique error structure (likely derived from the usage of the Phusion polymerase during library enrichment) [59]. We generated eight additional call sets for the Saqqaq genome using combinations of the following strategies: (1) sampling one random allele at each site or calling a majority rule consensus, (2) trimming five or zero nucleotides from the ends of the reads, and (3) including all sites or only those with a depth of coverage >5. In what follows we refer to the Saqqaq called genotypes as 'Saqqaq', and use a specific notation for the alternative versions (Figure S21). We computed *D*-statistics using both transversion polymorphisms only and all sites, in order to assess the potential bias introduced by typical aDNA C-T misincorporations. Note that this analysis was based on allele frequencies for which we pooled individuals coming from the same population.

17.3.1. Koryaks carry the largest East Asian ancestry proportion.

The model shown in Figure 3 indicates that Koryaks carry the largest proportion of East Asian ancestry among the modeled groups (including Native Americans), followed by the Saqqaq genome. We computed *D*-statistics of the form *D(H1, Koryaks; Han, Outgroup)* and *D(H1, Saqqaq; Han, Outgroup)*, where H1 includes all other admixed groups and the eight alternative Saqqaq callsets (Figure S21a). In agreement with the model, we found that Koryaks are significantly closer to Han than are the other admixed groups. However, we found Saqqaq (called genotypes) to be slightly closer to Han than Koryaks are (*D(Koryaks, Saqqaq; Han, Yoruba)=-0.006; Z=-1.02*). We observed that removing transition polymorphisms led to an increased value of *D*. Thus, we ruled out that this discrepancy was caused by additional, *post-mortem*-damage-derived, C-T substitutions in the Saqqaq genotypes. In that case, the expectation would be the opposite, as additional error in Saqqaq would result in artificial 'attraction' towards the outgroup [76]. On the other hand, alternative call sets conformed to the expectation of Koryaks being closer to the Han Chinese than Saqqaq is. We observed that all consensus-based call sets yielded variable results when filtering transition polymorphisms. Moreover, trimming five nucleotides from the ends of the reads did not affect *D* estimates notably. We expect the Phusion polymerase used for Saqqaq library enrichment not to incorporate C-T substitutions[59]; therefore, we consider the observed *D* patterns to be unrelated to *post-mortem*-damage-derived substitutions (expectedly located towards the ends of the reads [38]) in the Saqqaq data. In contrast, an overall higher error rate might lead to a mapping bias against reads containing authentic variants. Based on these results, we hypothesize that the unique error structure in Saqqaq resulted in biased genotype and consensus calls, which do not conform to the patterns expected for typical aDNA data [38], especially for comparisons with closely

related groups such as Koryaks. In what follows we only consider an alternative Saqqaq version for which we sampled random alleles, considered full-length reads and did not apply any depth filter (SaqqaqRnd). We deem this to be the one call set that yields the least biased estimates and incorporates the most data.

17.3.2. Bi-directional gene flow

The admixture graph approaches described above do not model admixture as a bi-directional process; instead, they both rely on directed edges. Based on these methods, we concluded that the most likely source for the additional Asian component in Athabascans is a population most closely related to Koryaks. Since a Native American admixture component has been documented in geographically close Siberian populations such as the Chukchis [60], we explored if Koryaks carry a similar genetic signature. We computed *D*-statistics of the form *D(Koryaks, Saqqaq; H3, Outgroup),* where *H3* represents the other non-East Asian groups (Figure S21b). We found Koryaks to be closer to Native Americans than Saqqaq is. While these results are consistent with bi-directional migration, gene flow between Koryaks and other Siberians populations that bear Native American ancestry and that are geographically closer to the New World, may also explain these results. Interestingly, USR1 did not yield *D* estimates as negative as those obtained for other Native Americans. In this case, *D* corresponds to the length of the branch that leads from the common ancestor of Koryaks, Saqqaq and USR1 on the one hand, to the common ancestor of USR1 and the introgressing Native American population on the other. In Sections 18 and 19, we find that the USR1 ancestors diverged from other Native Americans as early as ~20 kya, soon after their divergence from East Asians. Therefore, we expect a lower *D* estimate for USR1, as a consequence of such branch being shorter than that giving rise to NNA and SNA. Note that this bi-directional admixture event is not explicitly modeled in Sections 18 and 19. The shallow divergence time between Koryaks and Native Americans, the small admixture proportion in Athabascans (and potentially in Koryaks), and the unavailability of genomic data from multiple individuals make it hard to substantiate this result using demographic inference methods such as *diCal2* (Section 18) and *momi2* (Section 19). However, due to the expected low admixture proportions, we consider that bi-directional admixture will not bias our main findings considerably.

17.3.3. USR1 forms a clade with Native Americans.

Results based on admixture graph modeling support other analyses showing that USR1 forms a clade with Native Americans, to the exclusion of Old World populations (Sections 9, 10.4, 10.5, 14, 15, 16, 19). We confirmed these results by computing *D(USR1, H2; H3, Outgroup)*, where *H2* represents Native Americans and *H3* represents Old World populations (Figure S22a). As expected, we did not observe significant deviations from *D=0,* except for Athabascans, which bear Asian-related admixture. Note that we observed a non-significant systematic bias towards Native Americans being closer to non-Africans than USR1 is. We explored this pattern in Sections 11, 12

and 13 and concluded that it is a consequence of slightly increased Denisovan-related ancestry in USR1, which did not enter her ancestors through an independent admixture event. Rather, it is most likely a product of random sampling. Interestingly, we found that this shift was reduced for tests that involved Anzick1 and UstIshim, possibly as a consequence of considering three aDNA samples jointly in the *D*-statistic computation.

17.3.4. Koryaks are the best proxy for Asian ancestry in Athabascans.

We found that Athabascans are more closely related to Koryaks than to other Eurasian populations. We confirmed these results by computing *D(Athabascans, H2; H3, Outgroup)* where *H2* represents other Native Americans and *H3* represents Eurasian populations (Figure S22b). In agreement with the *treemix* and *qpgraph* results, Koryaks yielded the lowest *D*-statistics. Due to the early divergence between USR1 and other Native Americans (Sections 18 and 19), the test involving USR1 is particularly informative when assessing if there has been direct or indirect gene flow from Native Americans into Koryaks (Section 17.3.2). In this case, USR1 would likely represent an outgroup to Native Americans, including the 'introgressing' population. Indeed, we observed that *D* reaches its lowest values when USR1 is one of the ingroups.

17.3.5. Greenlandic Inuit bear Native American ancestry that is not present in Saqqaq.

Our model suggests that even though the Palaeo-Eskimo expansion had reached Greenland ~5kya [59], the Saqqaq genome does not bear Native American ancestry, as opposed to contemporary Inuit, who reached the same region during the last two millennia [71] and carry ~30% Native American ancestry. We tested this result by computing *D(Greenlanders, Saqqaq; H3, Outgroup)* where *H3* represents Native Americans (Figure S23b). As expected, we were able to reject the null hypothesis of Inuit and Saqqaq being equally related to Native Americans. Instead, we found support for gene flow with Native Americans. Again, we found that USR1 yielded *D* values closer to zero than other Native Americans. Similar to Section 17.3.2, we interpret this as being a consequence of the deep divergence between USR1 and other Native Americans.

17.3.6. Robustness to genotyping errors in Saqqaq

Our assessment suggests that the most problematic issue in our procedure is that the Saqqaq called genotypes have an error profile that we cannot account for by removing transition polymorphisms or calling a consensus. We tested the robustness of our model by repeating the *qpGraph* fitting approach (Section 17.2) using a new dataset where we substituted Saqqaq genotypes for a randomly sampled allele (SaqqaqRnd) (Figure S24). We obtained an admixture graph that produced a better fit (*maximum Z=3.048; p-val > 0.002*), than that observed for the original dataset (*maximum Z=3.269; p-val > 0.001*). Moreover, the main conclusions regarding the origins of Athabascans and Inuit did not

change. However, we observed discrepancies in the subgraph involving the ANE component in Koryaks, Eskimos and Native Americans. Given that we lack additional data points that would inform the inference of this subgraph (MalTa is the only ANE genome available to date [66]), we conclude that this section of the graph cannot be estimated with confidence.

17.4. A time frame for Asian introgression into Athabascans

Dating the admixture event that gave rise to most contemporary Na-Dene-speakers, is challenging due to the low degree of differentiation between Asians and Native Americans [89], as well as the reduced availability of genome-wide data from Na-Dene-speakers that do not carry post-Columbian admixture [31,60,61]. However, ancient genomic data is helpful in establishing temporal bounds for such event. While the population represented by USR1 most likely occupied interior Alaska at least until 11.5 kya [74] (the age of USR1), it is present-day Na-Dene-speakers who presently occupy the region [95]. Coupled with the geographic distribution of the Na-Dene language family, results showing that USR1 does not carry the Asian-derived ancestry component indicate that such ancestry entered the Americas after 11.5 kya, the time in which USR1 lived.

We used low-depth genomic data from three individuals (termed 939, 302 and 443) (Table S3) that lived in the Pacific Northwest Coast ~6-1.7 kya, to establish a lower bound for the date of this admixture event [31,100]. Note that hese individuals have been previously shown to be most closely related to contemporary Na-Dene-speaking groups from the Pacific Northwest Coast region [100]. For each individual, we fitted two separate admixture graphs similar to that shown in Figure 3, but excluding the Greenlandic Inuit. We substituted the two high-coverage contemporary Athabascan genomes with the ancient individual and fitted 1. a model in which the ancient individual is modeled as a 'non-admixed' NNA and 2. a model including the admixture event shown in Figure 3 for Athabascans. For both individuals 939 and 302, the 'non-admixed' model was rejected but the 'admixed' model yielded a maximum $|Z|$=2.8 and 2.5, respectively. Moreover, the 'admixed' model recapitulated the admixture proportions inferred for contemporary Athabascans (~90%) Figure 3. In contrast, the 'admixed' model for individual 443 yielded a maximum $|Z|$=3.28. Altogether, these results indicate that the admixture event that gave rise to most Na-Dene-speakers, between NNA and a Siberian population occurred well after 11.5 kya and at least prior to ~2.5 kya (the age of individual 302).

18. Demographic inference using dical2

We analyzed a subset of the whole-genome sequence data (Section 6.2) using the method *diCal 2.0* [101], which was previously employed in a related study [31]. Briefly, this method is based on the sequentially Markov coalescent approximation [102–104], and it uses a composite likelihood approach to estimate demographic parameters. The composite likelihood is obtained by multiplying demography-dependent conditional sampling probabilities [105,106], each of which corresponds to the probability of sampling a

new haplotype, conditional on having already observed a set of haplotypes. We focused on analyzing the genomic data from the Athabascan, the Karitiana, the Aymara, the Nivkh, the Koryak, and the Han Chinese in relation to the USR1 individual (an analysis of these populations not including USR1 was conducted in [31]). Our goal was to understand the key demographic events relating these populations.

18.1 Estimation of population size history

We first used *diCal 2.0* to estimate population size histories. To this end, we used the first 5 chromosomes and analyzed each individual separately. We divided the population size history into 6 epochs of constant population size (with boundaries at 0 kya, 10 kya, 20 kya, 40 kya, 60 kya, 70 kya, and ∞), and estimated the respective population sizes (Figure S25). Note that we shifted the estimates for USR1 by 11.5 ky, which is the estimated age of this ancient sample.

We observed good agreement between the population size histories estimated from different individuals of the same population. Furthermore, the ancestral population size prior to 70 kya was consistently estimated to be ~20,000 in all populations. Between 20 kya and 70 kya, we observed a strong bottleneck of roughly equal size in all populations. For times more recent than 20 kya, the estimated population sizes show larger differences across populations.

In Section S7 of [31], the population size before 70 kya N_A was set to 20,000, while the bottleneck size N_B, between N_A and the population divergence time T_{DIV} (on the order of 20 kya) was set to 1,800. However, we found N_B to be correlated with T_{DIV}, so we did not fix N_B in the analysis presented here. For more recent population sizes, we found the results reported in [31] to be in reasonable agreement with the ones obtained here; therefore, for reasons of consistency, we used those previously reported sizes in our subsequent analysis. Nevertheless, [31] did not report recent sizes for Han Chinese, Aymara, and USR1, so we set those to the harmonic mean of the estimated sizes for the two most recent epochs (Figure S25). In summary, for the subsequent analysis, we set the ancestral size N_A and the extant population sizes to the following:

$N_A = 20,000$
$N_{USR1} = 2,875$
$N_{Athabascan} = 2,600$
$N_{Karitiana} = 1,650$
$N_{Aymara} = 1,530$
$N_{Koryak} = 3,200$
$N_{Nivkh} = 4,650$
$N_{Han} = 12,000.$

18.2 Pairwise estimation of population divergence patterns

Next, we analyzed pairs of populations including USR1 (dated to 11.5 kya) and each of the other populations in the dataset. *diCal 2.0* as described in [101] can be used to analyze extant samples, and thus it cannot be applied right away to include a sample of ancient DNA. We therefore developed an extension to the method to incorporate ancient samples into the demographic inference. Using this modified version of *diCal 2.0*, we could infer demographic parameters using USR1 together with extant samples.

We used 2 haplotypes from the USR1 individual sampled at 11,500 years before the present and analyzed them jointly with 2 haplotypes each from the Athabascan, the Karitiana, the Aymara, the Koryak, the Nivkh, and the Han Chinese populations, one at a time. Due to computational limitations, we used the first ten chromosomes from each of these 4 haplotpyes (accounting for ~2 Gb). We fixed T_B = 70 kya, the ancestral size N_A = 20, 000, and the extant population sizes to the values obtained in the previous section. We analyzed these population pairs under different models including a clean split (Figure S26a with $m = 0$ and no T_M), isolation with migration until the present (Figure S26a with $m \geq 0$ and $T_M = 0$), isolation with migration with a stopping time (Figure S26a) and isolation with migration with a stopping time and a second contact (Figure S26b); and we show maximum (composite) likelihood estimates in Tables S4, S5, S6 and S7 for each model, respectively, where the migration rate m is given in units of per-individual per-generation probability of migrating.

We found that when USR1 was paired with the Han Chinese or a Siberian genome, the gene flow rate m_2 was estimated to be very high, suggesting that the founding population ancestral to USR1 and Siberians had a very weak population structure for over 10 ky, starting around 35 kya. Further, our estimate of the stopping time T_M for such strong gene flow was around 24.5 kya for the USR1 and Koryak pair, which can be interpreted as the time when the USR1 population began to diverge substantially from Siberians. The first three rows of Table S7 suggest that USR subsequently diverged from other Native Americans between 20 to 23 kya. Interestingly, our analysis shows that gene flow between USR1 and SNA stopped ~ 10.5 kya, while genetic exchange between USR and NNA continued until much more recently.

18.3 Simulation study

18.3.1. Population size history

We performed a simulation study to examine the uncertainty in our estimated population size histories for the real data. We considered a model with 5 epochs (0 kya, 10 kya, 20 kya, 40kya, 70 kya, ∞) of constant population sizes (5,000, 20,000, 2,000, 2,500, 20,000), and simulated 16 datasets, each comprising 2 haplotypes with 5 chromosomes of size 200 Mb, resulting in 1 Gb of sequence data for each haplotype. We then applied *diCal 2.0* to estimate the population sizes from these simulated datasets (Figure S27a). We found that for times larger than 20 kya, population sizes were estimated with high accuracy. In addition, the strong fluctuation in times more recent

than 20 kya is averaged out over the two most recent epochs, and also shows only moderate variability.

18.3.2. Divergence times

To assess the accuracy of our divergence time estimates, we conducted an extensive simulation study, using parameter settings close to those obtained for the real data. Unless otherwise noted, each simulated dataset consist of 4 haplotypes, 2 in the first population sampled at present and two in the second population, sampled at 11.5 kya (like the USR1 individual). For each haplotype, we simulated 4 chromosomes of length 250 Mbp, resulting in 1 Gbp per haplotype. Furthermore, we fixed the following parameters for the simulation, as well as for the analysis of the simulated data:

$T_B = 70{,}000$ *kya*
$N_A = 20{,}000$
$N_1 = N_2 = 3{,}000$.

We estimated demographic parameters based on a clean split model for the data simulated under a clean split model with divergence time $T_{DIV} = 20$ kya and bottleneck size $N_B = 1, 800$. Table S8 shows that T_{DIV} was accurately estimated (standard deviation of 320 years) with only a slight downward bias (370 years). Moreover, N_B was consistently biased downward by 350, but showed very little variation. Similarly, inference under the 'isolation with migration' model for data simulated using a model with $T_{DIV} = 20$ kya, $N_B = 1{,}800$, and $m = 0.0001$ showed very little variability in the parameter estimates (Table S9). Again, the estimates are almost unbiased, except N_B, which shows a similar downward bias as before.

We performed a similar assessment for the 'isolation with migration' model where migration stops at time T_M. In all cases we set $T_{DIV} = 20$ kya and $N_B = 1{,}800$ and explored four different scenarios that correspond to different configurations of T_M and m: $T_M = 13$ kya and $m = 0.001$ (Table S10), $T_M = 13$ kya and $m = 0.0001$ (Table S11), $T_M = 6$ kya and $m = 0.0001$ (Table S12), and $T_M = 6$ kya and $m = 0.001$ (Table S13). In the latter, we fixed the migration rate for the analysis. Similar to the results for the two previous models, we observed that T_{DIV} and N_B were estimated with very little variability, while N_B was biased downward. However, the estimates of T_M and m appeared to be correlated and strongly biased. Thus, we conclude that estimating the exact details of the recent migration history with high confidence is not feasible, unless the migration rate m is fixed.

Finally, we carried out a similar analysis for the second contact model (Figure 25b). We simulated data under such model with the following parameters: $T_{DIV} = 35$ kya, $T_M = 25$ kya, $N_B = 1, 800$, $m_1 = 0.001$, and $m_2 = 0.01$, but considered two inference strategies; one in which we estimated these parameters jointly (Table S14), and one in which we fixed the migration rates and estimated the rest of the parameters (Table S15). Once

again, we found little variability in the estimates of T_{DIV} and N_B, with a downward bias in N_B. However, we observed again that the parameters describing the details of the recent migration patterns are correlated and biased. Thus, we conclude that these details cannot be estimated with strong confidence, unless the migration rates are fixed.

18.4 Testing for gene flow

Lastly, we explored if a clean split could be rejected in favor of an 'isolation with migration' model. Given that the *diCal 2.0* likelihood function is a composite likelihood, we cannot test two alternative hypotheses using a regular likelihood-ratio test. Thus, we determined a rejection threshold empirically, using a simulation study. We simulated data under a clean split model with T_{DIV} = 20 kya and estimated the demographic parameters under a clean split model (Table S16) and an 'isolation with migration' model (Table S17). Using all 20 simulated datasets, we estimated the distribution of the log-likelihood difference between both models, for which we found the 95th percentile to be ~15 (Figure S27b). Thus, a log-likelihood difference greater than 15 would lead to a rejection of the clean split model, at a significance level of 5%. To better understand the log-likelihood differences observed in real data, we also simulated data under an isolation with migration model (T_{DIV} = 20 kya, m = 0.0005), and analyzed the data under a clean split model (Table S18) and an 'isolation with migration' model (Table S19). As expected, we observed a significant increase in likelihood when using the 'isolation with migration' model (log-likelihood difference >52).

Based on the real data, we found that that log-likelihood differences between the clean split and 'isolation with migration' models (Tables S4, S5, and S20) were above the empirically determined threshold (~15). Thus, for all population pairs, we reject a model of divergence without subsequent gene flow at a significance level of 5%. Furthermore, Tables S6 and S7 show that the second contact model yields a much higher likelihood for the USR1-Nivkh and USR1-Han pairs, while there is not such a strong improvement for the USR1-Koryak pair, although the likelihood increases. This suggests that the second contact model describes the data better than a model without second contact. On the other hand, for the USR1-Native American pairs, we found no improvement in likelihood. Therefore, the model with gene flow stopping at T_M and no second contact thereafter describes the latter pairs best.

In conclusion, the results of the simulation studies show that T_{DIV} and N_B are estimated with little variability, and only some downward bias in N_B. However, the details of the more recent migration events are estimated with less confidence. Thus, we conclude that the USR1 divergence from the Asian populations started at T_{DIV} ~ 36 kya after which very strong gene flow (O(m) = 0.01) ensued; the diverging populations are close to being panmictic. This gene flow stopped around ~24 kya, which is consistent with previous estimates (Table S7 in [31]). In relation to the Native American populations, USR1 started diverging ~20 kya with subsequent gene flow stopping ~11 kya and no second contact migration (Table S6). However, for the Athabascan population, weaker

gene flow appears to continue for a more extended period, which fits with a more recent movement of Athabascans into Alaska (Sections S20 and S21).

18.5 Confidence intervals

We followed a parametric bootstrap strategy to obtain confidence intervals for T_M and T_{DIV}. In Section 18.4, we inferred that the split between USR1 and the Old World genomes in our analysis is best explained by the 'second contact' model (Figure S26b, Table S7), whereas the split between USR1 and other Native American genomes is best explained by the 'isolation with migration' model with migration stopping at T_M (Figure S26a, Table S6). T_{DIV} confidence intervals for the Old World were based on the values shown in Table S14 and confidence intervals for Native Americans were based on Table S10. We obtained the most consistent estimates for T_M when we fixed the migration rates in the simulation study (Section 18.3.2). Therefore confidence intervals for T_M were based on the values shown in Tables S15 and S13, respectively. Note however, that estimates of T_M should be taken with caution as the estimates depend on a fixed value of m. In Table S21, we present bias-corrected estimates and 95% confidence intervals based on the simulated bootstrap replicates. In the main text, we use these confidence intervals for the Old World-Native American divergence estimates and those obtained in Section 19.3 for the Native American estimates.

19. Placement of USR1 using the SFS

We used a second demographic inference method, based on the site frequency spectrum (SFS), for exploring the demographic history of Native Americans with respect to Asians and Siberians, with a specific focus on the demographic history of USR1. We inferred the placement of the USR1 branch onto a backbone demography consisting of the Han Chinese, Koryak, Athabascan and Karitiana populations. The number of individuals used in this analysis are $n = 1$ for USR1 and Han Chinese, and $n = 2$ for the other three populations (Section 6). Throughout this analysis, we assumed a generation time of $g = 29$ years and a mutation rate of $\mu = 1.25\times10^{-8}$ bp/gen.

19.1 Marginal population size histories

We first estimated the marginal size histories of each population from the whole-genome sequence data using SMC++ (Figure 4b) [107]. As expected, we observed that the Han Chinese show more evidence of recent growth than the Siberian and Native American populations. Of note is that the uptick in USR1 growth right before the 11.5 kya sampling time could be caused by spurious mutations, perhaps because the sequences originate from aDNA. However, we did not observe an increase in singletons in USR1 (see below).

These estimates exhibit good agreement with those obtained using diCal (Figure S25). Both methods estimate the ancestral effective population size to be approximately 2 ×

10^4 for all populations, and a sharp bottleneck that occurred around 60kya. With the exception of the Han, all populations have effective population sizes in the low thousands in the very recent past. Finally, both methods detect a sharp, recent increase in the effective population size of the Han.

19.2 Demographic inference using the SFS

Next we used the SFS-based inference package *momi* [108], with extensions for migration described in [109], to estimate split times and migration events for these populations. The overall SFS for the sample is shown in Table S22.

19.2.1 Estimation of backbone demography

We first focused on estimating the 'backbone' demography consisting of all populations except USR1. The full space of possible demographies includes a tree whose topology and branch lengths are not known, as well as an unknown number of migration events/timings/intensities. This is a very difficult space in which to perform inference, and the current data do not seem to have enough power to do so. We worked around this issue by fixing the tree topology based on known archaeological evidence, as well as a few plausible migration events and their times. We estimated migration intensities and divergence times using maximum likelihood. Note that all migrations were modeled as discrete pulses. Therefore, a migration event of strength p at time t from population 1 into population 2 indicates that at time t, each lineage ancestral to population 2 moves independently with probability p into population 1 (looking backwards in time).
Writing $t_{Pop1,Pop2}$ for the estimated divergence time between populations *Pop1* and *Pop2*, the assumed migration events were:

1. Migration from Han into Athabascan shortly after the Athabascan-Karitiana split at time $t = 0.9\times t_{Athabascan,Karitiana}$, meant to capture the high fraction of Han ancestry found in Athabascan.
2. A single symmetric pulse between Athabascan and Karitiana halfway between their divergence time ($t = 0.5\times t_{Athabascan,Karitiana}$) and present.
3. Symmetric migration between Koryak and Han at time $t = 0.8\times t_{Han,Koryak}$.
4. Residual gene flow from Koryak into the Athabascan-Karitiana ancestor immediately after their divergence. We model this as a short series of pulses of equal intensity leading up to the Athabascan-Karitiana split.
5. A 'second contact' for Koryak into Athabascan 2 kya. (We also estimated such an event for Koryak -> Karitiana, but the estimated migration fraction was very low ($p = 10^{-6}$) indicating little support for this event in the data; thus, we removed it from subsequent fits.)

Our estimate of the backbone demography is shown in Figure S28.

19.2.2 USR1 join-on time estimates

We then considered the attachment point of the USR1 lineage onto the backbone demography, which we infer by using a similar SFS-based maximum likelihood approach. Figures 4b and S28 are heat maps of the likelihood of USR1 attaching to the demography at varying locations and time points. We considered the following three different scenarios concerning gene flow between USR1 and the other populations.

1. The first scenario is a "clean split" where no gene flow is assumed to have occurred between USR1 and any of the other populations (Figure S29a).
2. In the second scenario, we allowed for residual gene flow ($p = 10^{-3}$) between USR1 and its parent branch 1 ky after the split. As shown in Figure 4b, this did not significantly affect the likelihood.
3. Finally, we considered a scenario where, in addition to residual gene flow, there is gene flow from USR1 into Karitiana and Athabascan in the recent past, as would be the case if USR1 was a member of an Alaskan ghost population which still exists to date (Figure S29b).

In all cases, we found a most likely join-on time of 20.9 kya. Note that this estimate is the same for all three scenarios as we are using grid search with a spacing of 777 years to estimate the likelihood at different parts of the tree. Thus, each scenario is finding an MLE roughly between 21.7 and 20.1 kya. We can compare these estimates with the pairwise estimates of divergence and migration obtained using *diCal* (Tables S4-7). The estimated join-on time of 20.9 kya obtained using the SFS-based method is most comparable to the second-contact model (Table S7) from *diCal*, in which the most likely T_{DIV} between USR1 and Athabascan (Karitiana) is 19.8 (22.7) kya. Our estimate for the divergence between USR1 and Koryak (23.3kya) and Han (25.6kya) is considerably lower than the estimates of T_{DIV} obtained by *diCal*. However, we note that in these cases *diCal* estimated strong gene flow ($m = O(10^{-2})$) between these population pairs for about 10ky, which may be difficult to distinguish from panmixia in terms of the signature it leaves on the SFS even when allowing for a small number of pulse migrations. If we instead focus on T_M, the stopping time of gene flow, then we again see good agreement between our estimated pairwise divergence times for USR1 and Koryak/Han, and those that were estimated by *diCal* (24.5 and 26.5kya, respectively).

19.3 Bootstrap analysis

Finally, we assessed the robustness of our estimates by performing a parametric bootstrap analysis. We used the program *msprime* [84] to generate bootstrap replicates using the parameters inferred in Section 19.2. Each replicate consisted of 30 chromosomes of length 100Mb simulated independently under the structured coalescent. A total of 102 replicates were simulated, and the entire estimation pipeline (Section 19.2) was run on each replicate. We note that the topology of the underlying backbone demography (Figure S28) was assumed known for these simulations.

We summarize our results in Tables S22–S24. We found that the mean estimates for the divergence times agree fairly well with the true simulated values (Table S23). However, deeper splits exhibited greater variation. Interestingly, the estimates of migration rates were noisy (Table S24), likely due to small sample sizes and recent divergence times, which make migration rates difficult to infer using these data. Despite this variation, there was good agreement for the estimated tree topology. In fact, in 100% of the bootstrap replicates, the most likely tree topology had USR1 joining on to the branch ancestral to Athabascan and Karitiana, just below the divergence with Koryak. Statistics for the distance between the Koryak-(Athabascan,Karitiana) split and the most likely USR1 join-on time are shown in Table S25, and Figure S30 shows the distribution of likelihood surfaces across all simulations. In the latter, the warmer area concentrated near the Koryak-AK split represents the most likely placement of the USR1 individual onto the backbone demography.

Overall, these results are consistent with the inference based on haplotype data (Section 18). Both methods suggest that Native Americans became isolated from Asians and Siberians ~25 kya, after which the ancestors of USR1 started diverging from other Native Americans ~20 kya. Our *diCal* analysis further suggests that gene flow between USR1 and Native Amerticans was maintained at least until 11.5 kya (the age of USR1). Yet, we infer that this gene flow did not involve the already differentiated North and South Native American branches (here represented by Athabascans and Karitiana, respectively), which diverged from each other ~15 kya, as we found USR1 to be equally related to both branches (Section 10.6).

20. Assessment of the splits within ancestral Native American populations.

Our results show that the population represented by USR1 belongs to one of two basal branches of ancestral Native Americans. We termed the branch leading to USR1 ***Ancient Beringians (AB)*** and the branch leading to other Native Americans ***Northern and Southern Native Americans (NNA+SNA)***. We assessed five possible scenarios for where this basal AB and NNA+SNA split and the subsequent split between North Native Americans and South Native Americans may have taken place[20,31,110–112] (see Figure 2 for a schematic representation). The following parameters, derived from this study and previous work, apply to our assessment of these scenarios:

1. Based on genomic evidence from modern individuals and USR1, the NNA-SNA split has been estimated as occurring between 14,000 and 17,000 years ago (Section 19), and no later than 12,600 years ago [44].

2. USR1 is a Native American outgroup to both the NNA and SNA lineages. Therefore, the split within the source population of ancient Native Americans that led to USR1 on the one hand, and other Native Americans on the other, must predate the NNA-SNA split. We estimate this split to have occurred approximately 20 kya after which, there

was a period of gene flow between AB and NNA and between AB and SNA (Sections 18 and 19).

3. Currently, the earliest secure archaeological evidence for a human presence in eastern Beringia is the site of Swan Point, Alaska, which is radiocarbon dated to 14.2 kya [113]. Claims for an earlier presence, as at Bluefish Caves [114], based on the presence of apparently humanly-modified bones, are intriguing, but ultimately unconvincing: only 15 bones out of 36,000+ examined prove to have marks that may be a consequence of human action, but since those marks are also consistent with other, non-human, processes, the results are not unequivocal. Furthermore, the new dated specimens are not directly tied to the complicated, cryoturbated stratigraphy of the site, or with the few human-produced stone artifacts, which are similar to later, Denali-age, materials (i.e., microblade cores of the Campus type, not the older Yubetsu type). Given that, and that the new radiocarbon ages are scattered over 10,000 years, the Bluefish Caves evidence is not compelling.

4. Radiocarbon dating puts the age of USR1 at ~11,500 years BP[1,2].

Scenario 1: Divergence of AB and NNA+SNA took place in Eurasia, and the NNA–SNA split took place in North America, south of Beringia.

If the divergence between AB and NNA+SNA was a clean split, this scenario implies that these lineages moved independently from Eurasia, and that AB remained genetically isolated from the NNA+SNA lineage, between the time of their arrival in eastern Beringia and ~11.5 kya (the age of USR1). Alternatively, if there was gene flow between these two lineages, this scenario implies that they moved together from the Eurasia as a weakly structured population. We assessed these alternatives in Section 19 and found support for gene flow between the ancestors of USR1 and the ancestors of NNA and SNA. This scenario is most consistent with the archaeological record, which to date lacks secure evidence of human occupation in Beringia and the Americas at ~20.1 kya. The migrations into Beringia and the Americas may have been concurrent in different geographic areas or sequential through the same geographic areas.

Regardless of the nature of the USR1 split, **Scenario 1** posits that the NNA and SNA lineages diverged once they were beyond the continental ice sheets, and after they became isolated from the AB population. Although we found NNA and SNA to be equidistant to USR1 (Figures 1 and S7), diversity-based estimates such as the *D*-statistic have little resolution in finding asymmetric gene flow between these recently diverged populations. However, we consider a region south of eastern Beringia, where the ancestors of USR1 were not present, to be the most plausible location for the NNA-SNA split ~15 kya. Human presence in lower-latitudes has been securely dated to 14.6 kya and later [115,116], and glacial ice could have provided the barrier that led to genetic isolation from the remnant AB population in eastern Beringia. The location of this split could have been along migration corridors or in central North America.

Scenario 2: Divergence of AB and NNA+SNA took place in eastern Beringia, and the NNA–SNA split took place in North America, south of Beringia.

This scenario implies that the ancestral Native American population entered eastern Beringia then diverged into AB and the ancestral NNA+SNA line. If that divergence was a clean split, strong population structure in eastern Beringia is required between ~20 kya (the estimated time of the divergence within the ancestral Native American) and ~11.5 kya (the age of USR1). Strong population structure is not required if there was gene flow, for which we find evidence in Section 19.

Similar to **Scenario 1**, this model implies that NNA+SNA became isolated from the remainder of the ancestral Native American population, again possibly by moving beyond the glacial ice sheets, and then in turn split into those basal branches.

Scenario 3: Divergence of AB and NNA+SNA took place in Eurasia, and the NNA–SNA split took place in eastern Beringia.

This and **Scenario 1** have the same implications regarding the divergence of AB and NNA+SNA. However, it requires population structure in eastern Beringia, strong enough to have the ancestors of USR1, and those of the NNA and SNA inhabiting the same region, but with highly restricted gene flow between them for the period from ~15.7 kya (estimated time for the NNA–SNA split (Section 18)) and the time in which NNA and SNA moved south of eastern Beringia as one structured or two separate migrations. Moreover, although genetic sampling of modern and ancient populations in the region is sparse, members of the SNA branch have not been documented in regions that were once north of the glacial ice [60,61]. In the absence of such strong structure and the requirement for a more complex mode of migration south of Beringia, we deem this scenario less plausible than **1** and **2**. This conclusion is further supported by our finding of gene flow between AB and NNA+SNA after their split ~20 kya (Section 19).

Scenario 4: All three lineages split in eastern Beringia.

This scenario has the same implications as **Scenario 2** regarding the AB and NNA+SNA split, and the same implications as **Scenario 3** regarding the NNA–SNA split. In the absence of strong structure in eastern Beringia between ~15.7 kya (the estimated time for the NNA–SNA split (Section 18)) and the time in which NNA and SNA moved south of eastern Beringia as one structured or two separate migrations, we consider this model to be less plausible than **1** or **2**.

Scenario 5: All three lineages split in Eurasia.

This scenario is similar to **Scenario 4** in that both sets of splits occurred in the same region. It requires strong population structure, particularly after the NNA–SNA split,

and for that structure to have persisted across the Old World and into Eastern Beringia over many millennia. Given that we find evidence for gene flow between AB and NNA+SNA, we consider these conditions to be unlikely. For that reason, we find this model to be less plausible than Scenarios **1** and **2**.

Of the five scenarios, **Scenario 1** and **Scenario 2** are the ones most likely, and though each has strengths and weaknesses in regard to the extant genetic and archaeological evidence, we cannot fully eliminate one or the other.

Both scenarios are compatible with our genetic results, but obviously they differ in regard to where the ancestors of USR1 split from other Native Americans. Viewed strictly from the genetic evidence, **Scenario 1** is less parsimonious. We found evidence of continuous gene flow between the ancestors of USR1 and NNA+SNA (Section 18), which would have been facilitated if these populations were geographically close between ~20 and 11.5 kya. Moreover, if the ancestors of USR1 travelled separately into eastern Beringia, it would imply they (or the ancestors of the NNA+SNA line, if they stayed behind) remained in Eurasia for a time period during which they must have stayed isolated from other Asian/Siberian populations, since USR1 shows no evidence of gene flow with Siberians/Asians, and still forms a clade with Native Americans (Section 10.4). Thus, by this genetic evidence, **Scenario 2** is favored.

However, the earliest securely dated archaeological remains in Beringia only date to ~15-14 kya [73,74,114,117–122]. Given that the split between the ancestors of USR1 and NNA+SNA is estimated to have occurred ~21 kya, it seems unlikely it would have occurred in eastern Beringia. Additionally, this inference is based on the fact that the split occurred during the LGM, a cold period unlikely to be associated with expanding populations moving northward into Beringia. Since isolation of ancestral Native Americans from ancestral East Asians was maintained for thousands of years in Northeast Asia, these isolation mechanisms (e.g., different geographical locations or ecological adaptations) likely persisted through the LGM. On the basis of archaeological and paleoecological evidence (and consistency with the genetic evidence presented above) **Scenario 1** is favored. Ultimately, additional genetic and archaeological data will fully resolve the location of the initial AB and NNA+SNA split.

As for the other scenarios (**3** through **5**), in order to be compatible with the genetic results, there would have to have been extended periods of strong population structure, for which we do not see compelling genetic evidence, and which we deem less likely to have occurred between what are known archaeologically to be highly mobile populations [74,110,123].

Finally, we observe that members of the NNA (including Athabascans) later moved northward, which must have occurred sometime after 11,500 years ago, [112,124], ultimately replacing or absorbing the population to which USR1 belonged.

21. Extended discussion integrating the genetic results with current archaeological evidence

Our findings can be integrated with current knowledge of Siberian, Beringian, and Paleoindian archaeology [74,110,125]. This integrated model posits divergence of Native Americans from East Asians initiating at ~36 kya, but with both groups behaving as an almost panmictic population until ~24kya (Section 18, similar estimates have been obtained based on modern genome data [31]). Ancestral Native American populations were still likely in Northeast Asia during this period. Although direct evidence is lacking as to the mechanism(s) that led to the cessation of gene flow, we note that this period coincides with evidence for local depopulation in far Northeast Asia [126], but see [127]. Importantly, there is no unequivocal evidence for human presence in the Americas during this period [74,114].

Following the LGM, there is only one widespread cultural tradition in Northeast Asia: Diuktai Culture (Late Upper Paleolithic) [128]. This tradition expanded from central Siberia to the northeast between ~18-14 kya, and it is related to the technology used by the earliest Beringians at Swan Point, Alaska, dating to 14.2 kya [113]. We note that USR1 (a descendent of the AB population) was part of a cultural tradition (Denali complex) with technological connections with Diuktai Culture. The Diuktai culture itself was replaced in Siberia by the Sumnagin Culture after ~12.6 kya, possibly representing a different population [129].

The ancient Native American population initially resided in Northeast Asia. The basal split between AB and NNA+SNA began at ~20 kya and there continued to be gene flow between these groups until ~11 (Section 18). While NNA+SNA populations maintained gene flow with AB, dispersal of one or both populations into different or perhaps distant geographic areas, or ecological settings that led to the development of different economic adaptations, may have resulted in the cessation of gene flow prior to the NNA-SNA split. Separate locations in Siberian and Beringian interior riverine (AB) and Northern Pacific coastal areas (NNA+SNA), may have led to the divergence between these groups. Such regions have been recently suggested to have high levels of productivity during this period [73]. Alternatively, both branches (AB and NNA+SNA) may have arrived as distinct pulses into eastern Beringia, separate in time but with gene flow between them.

Archaeological evidence of the earliest human occupations south of the Laurentide Ice Sheet and Cordilleran Glacier Complexes points to migration out of eastern Beringia by ~15-14 kya. This correlates with the estimated divergence time for NNA and SNA at ~15 kya (Section 19), and certainly prior to the age of Anzick at 12.6 kya [31,44], followed by rapid diversification of multiple major lineages within North America [77]. This scenario is consistent with the apparently rapid expansion and settlement of ecologically diverse regions of the American continents south of the glacial ice [110]. That NNA and

SNA are symmetrical to USR1 strongly suggests that the NNA-SNA split occurred south of Beringia, with later gene flow with Na-Dene (NNA) ancestors who must have moved north after deglaciation.

Our data suggest that Na-Dene (Athabascans) would have to have diverged from the other NNA south of Beringia, and then moved northward at a later period. One hypothesis consistent with the genetic and archaeological data posits an AB population in interior Beringia, associated with the Denali complex (Figure S1) in the early Holocene with gene flow with Na-Dene ancestors until ~6-4 kya. The major cultural transition in this region occurs at ~6 kya, when the Denali complex is replaced by the Northern Archaic tradition [24]. While there is some material cultural continuity (*e.g.,* wedge-shaped microblade cores), new technologies enter the record that appear to relate to earlier forms in the western Subarctic to the south and east [18,24]. This may reflect population replacement or absorption.

Our data indicates more complex genetic history of Native American ancestors prior to their basal divergence into SNA and NNA branches ~15 kya that extends prior to their migration into the Americas. This divergence within the ancestral Native American population and later within NNA and SNA occurred between ~21-11 kya, and then again sometime after 11 kya when descendants of NNA moved north.

There are broader implications of our data and analyses for the nature, location and timing of the colonization of the Americas, which has been historically approached by using the Beringian Incubation or Standstill Model [52,130]. We found that gene flow between East Asians (Koryak, Nivkh) and Native Americans was interrupted 24 kya, after which NNA and SNA diverged ~15 kya south of Beringia. This implies a ~10,000 year period of isolation of ancestral Native Americans (in broad agreement with [31] and [77]), in contrast to the originally proposed 15,000 year estimate [52]. Moreover, this 'incubation' period is now shown to be more complex, marked early on by a single ancestral Native American population, which later split into two populations with ensuing gene flow (AB and NNA+SNA groups). Yet, these groups maintained genetic isolation from East Asian and Ancient North Eurasian groups. The geographic locations of these populations remain ambiguous and cannot be determined on the basis of genetic data alone. However, from an archaeological standpoint, the lack of evidence of human occupation of eastern Beringia or the Americas prior to ~15-14 kya suggests that the period of isolation between Asian and ancestral Native Americans occurred in Northeast Asia, where there is evidence of human occupations during this period (~24-14 kya) [127].

Recent genetic analyses have suggested some South Native American populations derive a fraction of their ancestry from ancestors with closer relationships to modern Australasians, termed 'Population Y' [68]. Two alternative scenarios are consistent with the data: (1) the Australasian signal came in during recent times, *i.e.*, during the Holocene, differentially reflected in a few populations [31], or (2) the Australasian signal

reflects a remnant of an earlier migration from a structured Beringian population that later admixed with newly arrived ancestral Native Americans in South America [68,131]. The latter would imply that 'Population Y' diverged from other Native Americans prior to the AB-NNA+SNA split, as we did not detect this Australasian-related signal in USR1. However this issue will only be resolved with geographically and temporally denser sampling across the Americas, particularly in Beringia and South America.

The relative genetic position of AB, within the broader Native American ancestral population, suggests long-term residence in Interior Eastern Beringia, from at least 14-11.5 ka, and perhaps extending to ~6 kya. This genetic continuity is congruent with evidence of technological continuity between the Eastern Beringian tradition (including the Chindadn complex) and the later Denali complex [21,74,113,132]. However, we are limited by the current Na-Dene genome samples. Na-Dene genomes from Alaska, where the linguistic diversity is greatest [94,98], could help resolve the ambiguities with USR1-Na-Dene relationships, and the later history of the relict Ancient Beringian population that USR1 represents.

References

1. Potter, B. A., Irish, J. D., Reuther, J. D., Gelvin-Reymiller, C. & Holliday, V. T. A Terminal Pleistocene Child Cremation and Residential Structure from Eastern Beringia. *Science* **331,** 1058–1062 (2011).
2. Potter, B. A., Irish, J. D., Reuther, J. D. & McKinney, H. J. New insights into Eastern Beringian mortuary behavior: A terminal Pleistocene double infant burial at Upward Sun River. *Proc. Natl. Acad. Sci.* **111,** 17060–17065 (2014).
3. Potter, B. A., Reuther, J. D., Bowers, P. M. & Gelvin-Reymiller, C. Little Delta Dune site: A late-Pleistocene multicomponent site in central Alaska. *Curr Res Pleistocene* **25,** 132–135 (2008).
4. Halffman, C. M. *et al.* Early human use of anadromous salmon in North America at 11,500 y ago. *Proc. Natl. Acad. Sci.* **112,** 12344–12348 (2015).
5. Liversidge, H. M. & Molleson, T. Variation in crown and root formation and eruption of human deciduous teeth. *Am. J. Phys. Anthropol.* **123,** 172–180 (2004).
6. AlQahtani, S. J. Atlas of human tooth development and eruption (Queen Mary and Westfield College). Available at: www.atlas.dentistry.qmul.ac.uk.
7. Schutkowski, H. Sex determination of infant and juvenile skeletons: I. Morphognostic features. *Am. J. Phys. Anthropol.* **90,** 199–205 (1993).
8. Turner II, C. G., Nichol, C. R. & Scott, G. R. Scoring procedures for key morphological traits of the permanent dentition: the Arizona State University dental anthropology system. in *Advances in dental anthropology* (eds. Kelley, M. A. & Larsen, C. S.) (Wiley-Liss, 1991).
9. Pinhasi, R. *et al.* Optimal Ancient DNA Yields from the Inner Ear Part of the Human Petrous Bone. *PLOS ONE* **10,** e0129102 (2015).
10. Gamba, C. *et al.* Genome flux and stasis in a five millennium transect of European prehistory. *Nat. Commun.* **5,** 5257 (2014).

11. Tackney, J. C. *et al.* Two contemporaneous mitogenomes from terminal Pleistocene burials in eastern Beringia. *Proc. Natl. Acad. Sci.* 201511903 (2015). doi:10.1073/pnas.1511903112
12. Powers, W. R., Guthrie, R. D. & Hoffecker, J. F. *Dry Creek: Archeology and paleoecology of a late Pleistocene Alaskan hunting camp*. (National Park Service, Washington D.C., 1983).
13. Hare, P. G., Hammer, T. J. & Gotthardt, R. M. in *Projectile point sequences in northwestern North America* (eds. Carlson, R. L., Magne, M. P. R. & Simon Fraser University) (Archaeology Press, Simon Fraser University, 2008).
14. Workman, W. B. *Prehistory of the Aishihik-Kluane Area, southwest Yukon Territory*. (1978).
15. Kunz, M., Bever, M. & Adkins, C. *The Mesa Site: Paleoindians above the Arctic Circle*. (BLM – Alaska Open File Report 86. U.S. Department of the Interior, Bureau of Land Management, Anchorage, 2003).
16. Rasic, J. Functional variability in the late Pleistocene archaeological record of eastern Beringia: a model of late Pleistocene land use and technology from northwest Alaska. in *From the Yenisei to the Yukon: interpreting lithic assemblage variability in late Pleistocene/early Holocene Beringia* (eds. Goebel, T. & Buvit, I.) (Texas A&M University Press, 2011).
17. Goebel, T. *et al.* Serpentine Hot Springs, Alaska: results of excavations and implications for the age and significance of northern fluted points. *J. Archaeol. Sci.* **40,** 4222–4233 (2013).
18. Potter, B. A. Radiocarbon chronology of central Alaska: Technological continuity and economic change. *Radiocarbon* **50,** 181 (2008).
19. Clark, D. W. & Gotthardt, R. M. Occasional Papers in Archaeology No. 6.
20. Dixon, E. J. Cultural chronology of central interior Alaska. *Arct. Anthropol.* 47–66 (1985).
21. West, F. H. The Archaeological Evidence. in *American beginnings: the prehistory and palaeoecology of Beringia* (University of Chicago Press, 1996).
22. Younie, A. M., Le Blanc, R. J. & Woywitka, R. J. Little Pond: A Microblade and Burin Site in Northeastern Alberta. *Arct. Anthropol.* **47,** 71–92 (2010).
23. Anderson, D. D. A Stone Age Campsite at the Gateway to America. *Sci. Am.* **218,** 24–33 (1968).
24. Esdale, J. A. A Current Synthesis of the Northern Archaic. *Arct. Anthropol.* **45,** 3–38 (2008).
25. Gilbert, M. T. P., Bandelt, H.-J., Hofreiter, M. & Barnes, I. Assessing ancient DNA studies. *Trends Ecol. Evol.* **20,** 541–544 (2005).
26. Willerslev, E. & Cooper, A. Review Paper. Ancient DNA. *Proc. R. Soc. B Biol. Sci.* **272,** 3–16 (2005).
27. Allentoft, M. E. *et al.* Population genomics of Bronze Age Eurasia. *Nature* **522,** 167–172 (2015).
28. Rohland, N. & Hofreiter, M. Ancient DNA extraction from bones and teeth. *Nat. Protoc.* **2,** 1756–1762 (2007).
29. Meyer, M. & Kircher, M. Illumina Sequencing Library Preparation for Highly

Multiplexed Target Capture and Sequencing. *Cold Spring Harb. Protoc.* **2010,** pdb.prot5448–pdb.prot5448 (2010).
30. Lindgreen, S. AdapterRemoval: Easy Cleaning of Next Generation Sequencing Reads. *BMC Res. Notes* **5,** 337 (2012).
31. Raghavan, M. *et al.* Genomic evidence for the Pleistocene and recent population history of Native Americans. *Science* **349,** aab3884–aab3884 (2015).
32. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25,** 1754–1760 (2009).
33. Schubert, M. *et al.* Improving ancient DNA read mapping against modern reference genomes. *BMC Genomics* **13,** 178 (2012).
34. DePristo, M. A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43,** 491–498 (2011).
35. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25,** 2078–2079 (2009).
36. 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).
37. Delaneau, O., Zagury, J.-F. & Marchini, J. Improved whole-chromosome phasing for disease and population genetic studies. *Nat. Methods* **10,** 5–6 (2012).
38. Briggs, A. W. *et al.* Patterns of damage in genomic DNA sequences from a Neandertal. *Proc. Natl. Acad. Sci.* **104,** 14616–14621 (2007).
39. Malaspinas, A.-S. *et al.* bammds: a tool for assessing the ancestry of low-depth whole-genome data using multidimensional scaling (MDS). *Bioinformatics* **30,** 2962–2964 (2014).
40. Orlando, L. *et al.* Recalibrating Equus evolution using the genome sequence of an early Middle Pleistocene horse. *Nature* **499,** 74–78 (2013).
41. Kent, W. J. *et al.* The Human Genome Browser at UCSC. *Genome Res.* **12,** 996–1006 (2002).
42. Korneliussen, T. S., Albrechtsen, A. & Nielsen, R. ANGSD: Analysis of Next Generation Sequencing Data. *BMC Bioinformatics* **15,** (2014).
43. Briggs, A. W. *et al.* Removal of deaminated cytosines and detection of in vivo methylation in ancient DNA. *Nucleic Acids Res.* **38,** e87–e87 (2010).
44. Rasmussen, M. *et al.* The genome of a Late Pleistocene human from a Clovis burial site in western Montana. *Nature* **506,** 225–229 (2014).
45. Fu, Q. *et al.* A Revised Timescale for Human Evolution Based on Ancient Mitochondrial Genomes. *Curr. Biol.* **23,** 553–559 (2013).
46. Green, R. E. *et al.* A Complete Neandertal Mitochondrial Genome Sequence Determined by High-Throughput Sequencing. *Cell* **134,** 416–426 (2008).
47. Calabrese, F. M., Simone, D. & Attimonelli, M. Primates and mouse NumtS in the UCSC Genome Browser. *BMC Bioinformatics* **13,** 1 (2012).
48. Katoh, K. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* **30,** 3059–3066 (2002).
49. Katoh, K. & Frith, M. C. Adding unaligned sequences into an existing alignment using MAFFT and LAST. *Bioinformatics* **28,** 3144–3146 (2012).
50. Gelman, A. & Rubin, D. B. Inference from iterative simulation using multiple

sequences. *Stat. Sci.* 457–472 (1992).
51. Vianello, D. *et al.* HAPLOFIND: A New Method for High-Throughput mtDNA Haplogroup Assignment. *Hum. Mutat.* **34,** 1189–1194 (2013).
52. Tamm, E. *et al.* Beringian Standstill and Spread of Native American Founders. *PLoS ONE* **2,** e829 (2007).
53. Skoglund, P., Storå, J., Götherström, A. & Jakobsson, M. Accurate sex identification of ancient human remains using DNA shotgun sequencing. *J. Archaeol. Sci.* **40,** 4477–4482 (2013).
54. Racimo, F., Renaud, G. & Slatkin, M. Joint Estimation of Contamination, Error and Demography for Nuclear DNA from Ancient Humans. *PLOS Genet.* **12,** e1005972 (2016).
55. Korneliussen, T. S. & Moltke, I. NgsRelate: a software tool for estimating pairwise relatedness from next-generation sequencing data. *Bioinformatics* btv509 (2015). doi:10.1093/bioinformatics/btv509
56. Albrechtsen, A. *et al.* Relatedness mapping and tracts of relatedness for genome-wide data in the presence of linkage disequilibrium. *Genet. Epidemiol.* **33,** 266–274 (2009).
57. Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* **19,** 1655–1664 (2009).
58. Purcell, S. *et al.* PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *Am. J. Hum. Genet.* **81,** 559–575 (2007).
59. Rasmussen, M. *et al.* Ancient human genome sequence of an extinct Palaeo-Eskimo. *Nature* **463,** 757–762 (2010).
60. Reich, D. *et al.* Reconstructing Native American population history. *Nature* **488,** 370–374 (2012).
61. Verdu, P. *et al.* Patterns of Admixture and Population Structure in Native Populations of Northwest North America. *PLoS Genet.* **10,** e1004530 (2014).
62. 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).
63. Greenberg, J. H. *et al.* The Settlement of the Americas: A Comparison of the Linguistic, Dental, and Genetic Evidence [and Comments and Reply]. *Curr. Anthropol.* **27,** 477–497 (1986).
64. Meyer, M. *et al.* A High-Coverage Genome Sequence from an Archaic Denisovan Individual. *Science* **338,** 222–226 (2012).
65. Prüfer, K. *et al.* The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature* **505,** 43–49 (2013).
66. Raghavan, M. *et al.* Upper Palaeolithic Siberian genome reveals dual ancestry of Native Americans. *Nature* **505,** 87–91 (2013).
67. Rasmussen, M. *et al.* The ancestry and affiliations of Kennewick Man. *Nature* (2015). doi:10.1038/nature14625
68. Skoglund, P. *et al.* Genetic evidence for two founding populations of the Americas. *Nature* (2015). doi:10.1038/nature14895
69. Fu, Q. *et al.* Genome sequence of a 45,000-year-old modern human from

western Siberia. *Nature* **514,** 445–449 (2014).
70. Li, J. Z. *et al.* Worldwide Human Relationships Inferred from Genome-Wide Patterns of Variation. *Science* **319,** 1100–1104 (2008).
71. Raghavan, M. *et al.* The genetic prehistory of the New World Arctic. *Science* **345,** 1255832–1255832 (2014).
72. Skotte, L., Korneliussen, T. S. & Albrechtsen, A. Estimating Individual Admixture Proportions from Next Generation Sequencing Data. *Genetics* **195,** 693–702 (2013).
73. Hoffecker, J. F., Elias, S. A., O'Rourke, D. H., Scott, G. R. & Bigelow, N. H. Beringia and the global dispersal of modern humans: Beringia and the Global Dispersal of Modern Humans. *Evol. Anthropol. Issues News Rev.* **25,** 64–78 (2016).
74. Potter, B. A., Holmes, C. E. & Yesner, D. R. Technology and Economy among the Earliest Prehistoric Foragers in Interior Eastern Beringia. in *Paleoamerican odyssey* (eds. Graf, K. E., Ketron, C. V. & Waters, M. R.) (Texas A&M University Press, 2014).
75. Patterson, N. *et al.* Ancient Admixture in Human History. *Genetics* **192,** 1065–1093 (2012).
76. Durand, E. Y., Patterson, N., Reich, D. & Slatkin, M. Testing for Ancient Admixture between Closely Related Populations. *Mol. Biol. Evol.* **28,** 2239–2252 (2011).
77. Llamas, B. *et al.* Ancient mitochondrial DNA provides high-resolution time scale of the peopling of the Americas. *Sci. Adv.* **2,** e1501385–e1501385 (2016).
78. Neves, W. A., Powell, J. F., Prous, A., Ozolins, E. G. & Blum, M. Lapa vermelha IV Hominid 1: morphological affinities of the earliest known American. *Genet. Mol. Biol.* **22,** 461–469 (1999).
79. González-José, R. *et al.* Craniometric evidence for Palaeoamerican survival in Baja California. *Nature* **425,** 62–65 (2003).
80. von Cramon-Taubadel, N., Strauss, A. & Hubbe, M. Evolutionary population history of early Paleoamerican cranial morphology. *Sci. Adv.* **3,** e1602289 (2017).
81. Pool, J. E. & Nielsen, R. Inference of Historical Changes in Migration Rate From the Lengths of Migrant Tracts. *Genetics* **181,** 711–719 (2008).
82. Seguin-Orlando, A. *et al.* Genomic structure in Europeans dating back at least 36,200 years. *Science* (2014). doi:10.1126/science.aaa0114
83. Auton, A. *et al.* A global reference for human genetic variation. *Nature* **526,** 68–74 (2015).
84. Kelleher, J., Etheridge, A. M. & McVean, G. Efficient coalescent simulation and genealogical analysis for large sample sizes. *PLoS Comput Biol* **12,** e1004842 (2016).
85. Sankararaman, S., Mallick, S., Patterson, N. & Reich, D. The Combined Landscape of Denisovan and Neanderthal Ancestry in Present-Day Humans. *Curr. Biol.* (2016). doi:10.1016/j.cub.2016.03.037
86. Fu, Q. *et al.* The genetic history of Ice Age Europe. *Nature* (2016). doi:10.1038/nature17993
87. Pickrell, J. K. & Pritchard, J. K. Inference of Population Splits and Mixtures from Genome-Wide Allele Frequency Data. *PLoS Genet.* **8,** e1002967 (2012).
88. Kimura, M. The number of heterozygous nucleotide sites maintained in a finite

population due to steady flux of mutations. *Genetics* **61,** 893 (1969).
89. Xing, J. *et al.* Toward a more uniform sampling of human genetic diversity: A survey of worldwide populations by high-density genotyping. *Genomics* **96,** 199–210 (2010).
90. Green, R. E. *et al.* A Draft Sequence of the Neandertal Genome. *Science* **328,** 710–722 (2010).
91. Gronau, I., Hubisz, M. J., Gulko, B., Danko, C. G. & Siepel, A. Bayesian inference of ancient human demography from individual genome sequences. *Nat. Genet.* **43,** 1031–1034 (2011).
92. Li, H. & Durbin, R. Inference of human population history from individual whole-genome sequences. *Nature* **475,** 493–496 (2011).
93. Scott, G. R. *et al.* Sinodonty, Sundadonty, and the Beringian Standstill model: Issues of timing and migrations into the New World. *Quat. Int.* (2016). doi:10.1016/j.quaint.2016.04.027
94. Kari, J. M. & Potter, B. A. *The Dene-Yeniseian connection*. (University of Alaska Department of Anthropology/Alaska Native Language Center, 2011).
95. Ruhlen, M. The origin of the Na-Dene. *Proc. Natl. Acad. Sci.* **95,** 13994–13996 (1998).
96. Flegontov, P. *et al.* Genomic study of the Ket: a Paleo-Eskimo-related ethnic group with significant ancient North Eurasian ancestry. *Sci. Rep.* **6,** 20768 (2016).
97. Flegontov, P. *et al. Na-Dene populations descend from the Paleo-Eskimo migration into America*. (2016).
98. Krauss, M. E. Na-Dene. *Curr. Trends Linguist.* **10,** 903–978 (1973).
99. Leppälä, K., Vendelbo Nielsen, S. & Mailund, T. admixturegraph: An R Package for Admixture Graph Manipulation and Fitting. *Bioinformatics* (2017). doi:10.1093/bioinformatics/btx048
100. Lindo, J. *et al.* Ancient individuals from the North American Northwest Coast reveal 10,000 years of regional genetic continuity. *bioRxiv* 093468 (2016).
101. Steinrücken, M., Kamm, J. A. & Song, Y. S. Inference of complex population histories using whole-genome sequences from multiple populations. *bioRxiv* (2015). doi:10.1101/093468
102. Marjoram, P. & Wall, J. D. Fast‘ coalescent’ simulation. *BMC Genet.* **7,** 16 (2006).
103. McVean, G. A. T. & Cardin, N. J. Approximating the coalescent with recombination. *Philos. Trans. R. Soc. B Biol. Sci.* **360,** 1387–1393 (2005).
104. Wiuf, C. & Hein, J. Recombination as a Point Process along Sequences. *Theor. Popul. Biol.* **55,** 248–259 (1999).
105. Sheehan, S., Harris, K. & Song, Y. S. Estimating Variable Effective Population Sizes from Multiple Genomes: A Sequentially Markov Conditional Sampling Distribution Approach. *Genetics* **194,** 647–662 (2013).
106. Steinrücken, M., Paul, J. S. & Song, Y. S. A sequentially Markov conditional sampling distribution for structured populations with migration and recombination. *Theor. Popul. Biol.* **87,** 51–61 (2013).
107. Terhorst, J., Kamm, J. A. & Song, Y. S. Robust and scalable inference of

population history from hundreds of unphased whole genomes. *Nat. Genet.* **49,** 303–309 (2016).
108. Kamm, J. A., Terhorst, J. & Song, Y. S. Efficient computation of the joint sample frequency spectra for multiple populations. *J. Comput. Graph. Stat.* **26,** 182–194 (2016).
109. Kamm, J. A. One and Two Locus Likelihoods Under Complex Demography. (University of California, Berkeley, 2015).
110. Meltzer, D. J. *First peoples in a new world: colonizing ice age America*. (University of California Press, 2009).
111. Munyikwa, K., Feathers, J. K., Rittenour, T. M. & Shrimpton, H. K. Constraining the Late Wisconsinan retreat of the Laurentide ice sheet from western Canada using luminescence ages from postglacial aeolian dunes. *Quat. Geochronol.* **6,** 407–422 (2011).
112. Pedersen, M. W. *et al.* Postglacial viability and colonization in North America's ice-free corridor. *Nature* (2016). doi:10.1038/nature19085
113. Holmes, C. E. The Beringian and Transitional Periods in Alaska. in *From the Yenisei to the Yukon: interpreting lithic assemblage variability in late Pleistocene/early Holocene Beringia* (eds. Goebel, T. & Buvit, I.) (Texas A&M University Press, 2011).
114. Bourgeon, L., Burke, A. & Higham, T. Earliest Human Presence in North America Dated to the Last Glacial Maximum: New Radiocarbon Dates from Bluefish Caves, Canada. *PLOS ONE* **12,** e0169486 (2017).
115. Gilbert, M. T. P. *et al.* DNA from Pre-Clovis Human Coprolites in Oregon, North America. *Science* **320,** 786–789 (2008).
116. Dillehay, T. D. *et al.* Monte Verde: seaweed, food, medicine, and the peopling of South America. *Science* **320,** 784–786 (2008).
117. Goebel, T., Waters, M. R. & O'Rourke, D. H. The late Pleistocene dispersal of modern humans in the Americas. *science* **319,** 1497–1502 (2008).
118. Elias, S. A., Short, S. K., Nelson, C. H. & Birks, H. H. Life and times of the Bering land bridge. *Nature* **382,** 60–63 (1996).
119. Meiri, M. *et al.* Faunal record identifies Bering isthmus conditions as constraint to end-Pleistocene migration to the New World. *Proc. R. Soc. B Biol. Sci.* **281,** 20132167–20132167 (2013).
120. Elias, S. & Crocker, B. The Bering Land Bridge: a moisture barrier to the dispersal of steppe–tundra biota? *Quat. Sci. Rev.* **27,** 2473–2483 (2008).
121. Pitulko, V. V. *et al.* The Yana RHS site: humans in the Arctic before the last glacial maximum. *Science* **303,** 52–56 (2004).
122. Pitulko, V. V., Pavlova, E. Y. & Basilyan, A. E. Mass accumulations of mammoth (mammoth 'graveyards') with indications of past human activity in the northern Yana-Indighirka lowland, Arctic Siberia. *Quat. Int.* **406,** 202–217 (2016).
123. Graf, K. E. Siberian Odyssey. in *Paleoamerican odyssey* (eds. Graf, K. E., Ketron, C. V. & Waters, M. R.) (Texas A&M University Press, 2014).
124. Ives, J. W., Froese, D., Supernant, K. & Yanicki, G. Vectors, Vestiges, and Valhallas - Rethinking the Corridor. in *Paleoamerican odyssey* (eds. Graf, K. E., Ketron, C. V. & Waters, M. R.) (Texas A&M University Press, 2014).

125. Hoffecker, J. F. & Elias, S. A. *Human ecology of Beringia*. (Columbia University Press, 2007).
126. Goebel, T. The 'microblade adaptation' and recolonization of Siberia during the late Upper Pleistocene. *Archeol. Pap. Am. Anthropol. Assoc.* **12,** 117–131 (2002).
127. Kuzmin, Y. V. Comments on Graf, Journal of Archaeological Science 36, 2009 ' "The Good, the Bad, and the Ugly": evaluating the radiocarbon chronology of the middle and late Upper Paleolithic in the Enisei River valley, south-central Siberia'. *J. Archaeol. Sci.* **36,** 2730–2733 (2009).
128. Vasil'ev, S. A. The Upper Palaeolithic of Northern Asia. *Curr. Anthropol.* **34,** 82–92 (1993).
129. Pitul'ko, V. Terminal Pleistocene—Early Holocene occupation in northeast Asia and the Zhokhov assemblage. *Quat. Sci. Rev.* **20,** 267–275 (2001).
130. Kitchen, A., Miyamoto, M. M. & Mulligan, C. J. A Three-Stage Colonization Model for the Peopling of the Americas. *PLoS ONE* **3,** e1596 (2008).
131. Skoglund, P. & Reich, D. A genomic view of the peopling of the Americas. *Curr. Opin. Genet. Dev.* **41,** 27–35 (2016).
132. Cook, J. P. The early prehistory of Healy Lake, Alaska. (University of Wisconsin-Madison, 1969).

Table S1. Mapping statistics for both USR individuals. Sequencing reads were processed as described in Section 3. We provide mapping statistics, overall error estimates (Section 4.2) and mtDNA-based contamination estimates (Section 4.4) for each sequencing library. USER-treated libraries are indicated by a 'u'.

Sample extract	Library	Total reads	Trimmed reads	Mapped	Mapped unique	% Duplication	% Endo	Avg. read Length	Avg. DoC	Avg. DoC MT	Avg. DoC X	Avg. DoC Y	% Error-rate	p(MT authentic) (95% CI)	C to T (1st 5' base)
USR1-E7	012293E7_L1	20283367	14258000	2191515	2149213	1.930	10.59	51.00	0	4.3	0	0	0.7135	0.993 (0.898-0.998)	0.142184
USR1-E8	012293E8_L1	19511449	15434664	8859928	8785820	0.836	45.02	58.31	0.2	15.1	0.2	0	0.8176	0.990 (0.963-0.999)	0.13094
USR1-E9	012293E9_L1	16735364	13110738	7293833	7238069	0.764	43.25	59.13	0.1	13.6	0.1	0	0.7062	0.998 (0.975-0.999)	0.134662
USR1-E10	012293E10_L1	13594823	10638189	3374268	3348821	0.754	24.63	56.07	0.1	6	0.1	0	0.7606	0.997 (0.956-0.999)	0.133338
USR1-E11	012293E11_L1	16534726	12671272	1222766	1212574	0.833	7.33	52.19	0	2.6	0	0	0.5369	0.990 (0.782-0.998)	0.109447
USR1-E7	12293E7u_L12	3498986470	2599477058	1005165724	647465313	35.586	18.50	51.52	10.8	135.4	11.2	0.1	0.0879	0.999 (0.996-0.999)	
USR1-E7	12293E7u_L13	1360860967	988326922	353640339	295306991	16.495	21.70	50.96	4.9	103.5	5	0	0.0887	0.999 (0.996-0.999)	
USR1-E8	012293E8_L2u	88948876	67867703	11547813	11109122	3.798	12.48	52.12	0.2	20.5	0.2	0	0.0991	0.998 (0.989-0.999)	
USR1-E9	012293E9_L2u	60666839	47427326	28805014	28248325	1.932	46.56	55.28	0.5	34.2	0.5	0	0.118	0.999 (0.995-0.999)	
USR1-E10	012293E10_L2u	7835321	6251576	3729794	3695474	0.920	47.16	57.91	0.1	6.9	0.1	0	0.1124	0.999 (0.976-0.999)	
USR1-E11	012293E11_L2u	94144930	73518825	25141269	24480289	2.629	26.00	53.27	0.4	32.4	0.4	0	0.1196	0.996 (0.989-0.999)	
	USR1	5111443403	3782869410	1428029953	1010305514	29.251	19.76	51.53	16.8	332.9	17.4	0.1	0.0899	0.998 (0.995-0.999)	
USR2-E2	12294E2_L2	324445067	150551141	6339546	500128	92.110	0.15	47.55	<0.1	<0.1	<0.1	-	0.3561	-	0.116875
USR2-E2	012294E2_L3u	44954160	27410188	1884500	1874437	0.533	4.16	47.88	<0.1	4.8	<0.1	-	0.0988	0.998 (0.964-0.999)	
USR2-E2	12294_E2_L4u	186043218	115855796	7088893	5068971	28.494	2.72	47.88	0.1	12.4	0.1	-	0.1167	0.999 (0.990-0.999)	
	USR2	230997378	143265984	8973393	6943408	22.622	3.00	47.88	0.1	17.2	0.1		0.1118	0.999 (0.988-0.999)	

Table S2. Relatedness between both USR individuals. We show the probabilities of sharing 0, 1 or 2 alleles identical by descent between pairs of genomes, as estimated by the two different approaches described in Section 5. For each comparison, the number of SNPs is shown in the rightmost column.

***NGSrelate* estimates**					
Individual 1	**Individual 2**	**k_0**	**k_1**	**k_2**	**#SNPs**
USR2	USR1	0.52	0.381	0.099	9811
USR2	Anzick	0.866	0.096	0.039	9525
USR1	Anzick	0.844	0.146	0.01	1259773
***relate* estimates**					
Individual 1	**Individual 2**	**k_0**	**k_1**	**k_2**	**#SNPs**
USR2	USR1	0.654	0.346	0	827
USR2	Anzick	1	0	0	809
USR1	Anzick	0.833	0.167	0	14112

Table S3. Whole genome sequencing reference dataset. Whole genomes used in this study are detailed. For each genome, we show the average depth of coverage, the number of sites with a called-genotype and a heterozygosity estimate based on called genotypes (Section 6.2).

Sample name	Name in this study	Population	Region	Avg. DoC	Non-missing genotypes	%Het	Ref.	Note
USR1	USR1	-	America	16.8	2436817386	0.05732	This st.	Ancient
USR2	USR2	-	America	0.1	-	-	This st.	Ancient
Anzick1	Anzick1	-	America	12.8	1608438381	0.05774	44	Ancient
Kennewick	Kennewick	-	America	1	-	-	67	Ancient
939	939	-	America	0.4	-	-	31	Ancient
Saqqaq	Saqqaq	Saqqaq	Greenland	14.4	1006515737	0.03994	59	Ancient
Mat'ta	Mal'ta	ANE	Siberia	1	-	-	66	Ancient
Denisovan	Denisovan	Denisovan	Altai	24.3	2500676863	0.01920	64	Archaic
Altai Neandertal	Neandertal	Neandertal	Altai	40.8	2524819423	0.01764	65	Archaic
DNK02	Dinka	Dinka	Africa	24.3	2569052451	0.09260	64	
HGDP00456	Mbuti	Mbuti	Africa	20.3	2548699461	0.09314	64	
HGDP00927	Yoruba	Yoruba	Africa	26.7	2567696800	0.09462	64	
HGDP01029	San	San	Africa	26.9	2568188751	0.09910	64	
HGDP01284	Mandenka	Mandenka	Africa	20.6	2558239985	0.09393	64	
Aleutian	Aleutian	Aleutian	America	20.8	2434924158	0.07021	71	95.03%masked
Athabascan_1	Athabascan1	Athabascan	America	23.2	2049773754	0.05848	71	
Athabascan_2	Athabascan2	Athabascan	America	22	2326962236	0.05735	71	
BI16	Karitiana2	Karitiana	America	23.4	2414562015	0.05214	44	
CEPH_11_D12	Pima	Pima	America	19	2252556878	0.04740	31	
HGDP00998	Karitiana	Karitiana	America	21.3	2565672745	0.05096	64	
HUI03	Huichol	Huichol	America	23.2	2340026621	0.05486	31	
SS6004476	Karitiana3	Karitiana	America	32	2642336844	0.05131	65	
SS6004479	Mixe	Mixe	America	36.3	2651934284	0.05711	65	
TA6	Aymara	Aymara	America	20.6	2304183514	0.05293	31	
Y2040	Yukpa	Yukpa	America	22.5	2232969390	0.05158	31	
mixe0002	MixeE	Mixe	America	41.6	2639493094	0.08888	68	

mixe0042	MixeF	Mixe	America	41.4	2639356258	0.08786	68	
HGDP00846	SuruiA	Surui	America	37.1	2639873887	0.07965	68	
HGDP00852	SuruiB	Surui	America	31.7	2639354004	0.07815	68	
HGDP00778	Han	Han	East Asia	22.3	2570665836	0.06727	64	
HGDP01307	Dai	Dai	East Asia	23.8	2562941592	0.06725	64	
HGDP00521	French	French	Europe	22.6	2564841578	0.07107	64	
HGDP00665	Sardinian	Sardinian	Europe	19.9	2560254118	0.06949	64	
Greenlander_1	Greenlander1	Greenlander	Greenland	44.2	2547084054	0.06227	71	
Greenlander_2	Greenlander2	Greenlander	Greenland	42.2	2541111608	0.06303	71	
HGDP00542	Papuan	Papuan	Oceania	21.6	2561366658	0.05695	64	
13748_3/HGDP00543	Papuan1	Papuan	Oceania	18.9	2564338201	0.05564	31	
13748_1/HGDP00541	Papuan2	Papuan	Oceania	18.6	2564785772	0.05603	31	
13748_7/HGDP00547	Papuan3	Papuan	Oceania	18.5	2569623095	0.05683	31	
13748_6/HGDP00546	Papuan4	Papuan	Oceania	18.3	2562414766	0.05531	31	
Bur1	Buryat1	Buryat	Siberia	21.4	2458344356	0.06579	31	
Bur2	Buryat2	Buryat	Siberia	25.8	2445545066	0.06792	31	
Ket1	Ket1	Ket	Siberia	34.5	1860142338	0.06879	31	
Ket2	Ket2	Ket	Siberia	23.5	2377467697	0.06466	31	
Kor1	Koryak1	Koryak	Siberia	21.4	2454102221	0.05793	31	
Kor2	Koryak2	Koryak	Siberia	20	2436604546	0.05899	31	
Nivh1	Nivkh1	Nivkh	Siberia	20.9	2419085671	0.05895	71	
Nivh2	Nivkh2	Nivkh	Siberia	22.6	2455741789	0.06185	71	
Yak1	Yakut1	Yakut	Siberia	23.1	2349556086	0.06355	31	
Yak2	Yakut2	Yakut	Siberia	21.1	2397931959	0.06562	31	

Table S4: **Results for the clean split model.** Demographic parameters inferred by *diCal 2.0*, using a clean split model (Figure 19a with $m = 0$ and no T_M), for each pair of genomes involving USR1. Parameters with the 'mle' prefix were estimated and the other parameters were fixed.

Pop 1	Pop 2	mle T_{DIV}	N_1	N_2	mle N_B	N_A	logLike
Athabascan	USR1	15882	2600	2875	1701	20000	-8424573
Aymara	USR1	15528	1530	2875	1680	20000	-9398840
Karitiana	USR1	16136	1650	2875	1650	20000	-9770161
Koryak	USR1	24709	3200	2875	1649	20000	-10620881
Nivkh	USR1	26430	4650	2875	1635	20000	-10768921
Han	USR1	28460	12000	2875	1578	20000	-12142319

Table S5: Results for the 'isolation with migration' model. Demographic parameters inferred by *diCal 2.0*, using an IM model (Figure 19a with $m \geq 0$ and $T_M = 0$), for each pair of genomes involving USR1. Parameters with the 'mle' prefix were estimated and the other parameters were fixed.

Pop 1	Pop 2	mle T_{DIV}	mle m	N_1	N_2	mle N_B	N_A	logLike
Athabascan	USR1	19635	3.90E-04	2600	2875	1643	20000	-8424538
Aymara	USR1	20292	3.40E-04	1530	2875	1610	20000	-9398807
Karitiana	USR1	21725	3.50E-04	1650	2875	1564	20000	-9770015
Koryak	USR1	35662	4.90E-04	3200	2875	1375	20000	-10620414
Nivkh	USR1	35459	3.60E-04	4650	2875	1381	20000	-10768526
Han	USR1	34199	2.30E-04	12000	2875	1399	20000	-12142137

Table S6: Results for the 'isolation with migration' model with migration stop. Demographic parameters inferred by *diCal 2.0*, using an IM model, where gene flow stops at T_M (Figure 19a with $m \geq 0$ and $T_M \geq 0$), for each pair of genomes involving USR1. Parameters with the 'mle' prefix were estimated and the other parameters were fixed.

Pop 1	Pop 2	mle T_M	mle T_{DIV}	mle m	N_1	N_2	mle N_B	N_A	logLike
Athabascan	USR1	2401	19714	4.80E-04	2600	2875	1641	20000	-8424535
Aymara	USR1	11397	21274	1.40E-03	1530	2875	1610	20000	-9398755
Karitiana	USR1	10310	22702	1.00E-03	1650	2875	1556	20000	-9769984
Koryak	USR1	2344	35677	5.40E-04	3200	2875	1373	20000	-10620404
Nivkh	USR1	1816	35390	3.90E-04	4650	2875	1385	20000	-10768523
Han	USR1	22584	34251	9.50E-03	12000	2875	1399	20000	-12142101

Table S7: **Results for the second contact model.** Demographic parameters inferred by *diCal 2.0* using the second contact model depicted in Figure 19b, for each pair of genomes involving USR1. Parameters with the 'mle' prefix were estimated and the other parameters were fixed.

Pop 1	Pop 2	mle T_M	mle T_{DIV}	mle m_1	mle m_2	N_1	N_2	mle N_B	N_A	logLike
Athabascan	USR1	4071	19822	3.50E-04	4.10E-04	2600	2875	1638	20000	-8424537
Aymara	USR1	11362	21221	2.50E-08	1.40E-03	1530	2875	1612	20000	-9398755
Karitiana	USR1	10500	22729	2.50E-08	1.10E-03	1650	2875	1557	20000	-9769984
Koryak	USR1	24459	36194	2.50E-03	3.40E-02	3200	2875	1361	20000	-10620398
Nivkh	USR1	26151	36327	2.00E-03	3.30E-02	4650	2875	1360	20000	-10768482
Han	USR1	26459	35556	1.20E-03	2.70E-02	12000	2875	1362	20000	-12142027

Table S8: Inference results using the clean split model for simulated data. Demographic parameters inferred by *diCal 2.0* using the clean split model for data simulated under the same model (Figure S19a, m = 0, no T_M) with T_{DIV} = 20 kya. Parameters with the 'mle' prefix were estimated and the other parameters were fixed.

	mle T_{DIV}	N_1	N_2	mle N_B	N_A	logLike
Truth	**20000**	**3000**	**3000**	**1800**	**20000**	
sim1	19365			1421		-7751363
sim2	19367			1424		-7732439
sim3	20278			1468		-7803315
sim4	19867			1458		-7859154
sim5	19745			1430		-7787484
sim6	19521			1452		-7859325
sim7	19566			1428		-7886269
sim8	19363			1451		-7929714
Mean	19634			1441		
SD	320			18		

Table S9: Inference results using the 'isolation with migration' model for simulated data. Demographic parameters inferred by *diCal 2.0* using the IM model for data simulated under the same model (Figure S19a, $T_M = 0$) with $T_{DIV} = 20$ kya and $m = 0.0001$. Parameters with the 'mle' prefix were estimated and the other parameters were fixed.

	mle T_{DIV}	mle m	N_1	N_2	mle N_B	N_A	logLike
Truth	**20000**	**1.00E-04**	**3000**	**3000**	**1800**	**20000**	
sim1	19647	2.50E-04			1426		-7794316
sim2	20422	2.30E-04			1424		-7832268
sim3	19814	1.60E-04			1432		-7823589
sim4	19873	2.30E-04			1442		-7864893
sim5	20147	1.60E-04			1446		-7794492
sim6	19701	1.20E-04			1426		-7774435
sim7	20081	1.30E-04			1432		-7861377
sim8	20024	2.80E-04			1446		-7740625
Mean	19964	1.90E-04			1434		
SD	257	6.00E-05			9		

Table S10: Inference results using the 'isolation with migration' model where gene flow stops at T_M = 13 kya, for simulated data. Demographic parameters inferred by *diCal 2.0* using the IM model with gene flow stopping time (Figure S19a), for data simulated under the same model with $T_{DIV} = 20$ kya and $m = 0.001$. Parameters with the 'mle' prefix were estimated and the other parameters were fixed.

	mle T_M	mle T_{DIV}	mle m	N_1	N_2	mle N_B	N_A	logLike
Truth	**13000**	**20000**	**1.00E-03**	**3000**	**3000**	**1800**	**20000**	
sim1	11632	20320	1.60E-03			1406		-7841541
sim2	11014	20192	1.30E-03			1443		-7818743
sim3	10712	20389	8.00E-04			1413		-7780795
sim4	11006	19999	2.00E-03			1418		-7794456
sim5	10432	19967	1.40E-03			1423		-7762155
sim6	11600	19144	1.50E-03			1449		-7756813
sim7	11915	20290	1.10E-03			1476		-7917835
sim8	11612	20138	2.40E-03			1431		-7792671
Mean	11240	20055	1.50E-03			1432		
SD	523	397	5.00E-04			23		

Table S11: Inference results using the 'isolation with migration' model where gene flow stops at T_M = 13 kya, for simulated data, but with a lower *m*. Demographic parameters inferred by *diCal 2.0* using the IM model with gene flow stopping time(Figure S19a), for data simulated under the same model with T_{DIV} = 20 kya and m = 0.0001. Parameters with the 'mle' prefix were estimated and the other parameters were fixed.

	mle T_M	mle T_{DIV}	mle m	N_1	N_2	mle N_B	N_A	logLike
Truth	**13000**	**20000**	**1.00E-04**	**3000**	**3000**	**1800**	**20000**	
sim1	11606	20109	8.30E-04			1445		-7869217
sim2	16518	18771	8.40E-03			1498		-7817827
sim3	15380	18700	3.30E-04			1432		-7889281
sim4	16564	19464	3.20E-03			1430		-7699842
sim5	15134	18110	4.20E-05			1503		-7937893
sim6	16039	19127	3.60E-04			1468		-7792691
sim7	14358	19220	8.40E-04			1437		-7815753
sim8	16943	19018	2.80E-02			1446		-7806453
Mean	15318	19065	5.30E-03			1457		
SD	1727	587	9.70E-03			29		

Table S12: **Inference results using the 'isolation with migration' model where gene flow stops at T_M = 6 kya, for simulated data.** Demographic parameters inferred by *diCal 2.0* using the IM model with gene flow stopping time (Figure S19a), for data simulated under the same model with T_{DIV} = 20 kya and m = 0.0001. Parameters with the 'mle' prefix were estimated and the other parameters were fixed.

	mle T_M	mle T_{DIV}	mle m	N_1	N_2	mle N_B	N_A	logLike
Truth	**6000**	**20000**	**1.00E-04**	**3000**	**3000**	**1800**	**20000**	
sim1	11563	20234	9.20E-04			1444		-7933793
sim2	9302	20185	6.10E-04			1432		-7786055
sim3	9570	20872	8.50E-04			1434		-7823960
sim4	11561	20515	8.20E-04			1436		-7924189
sim5	11553	20475	6.50E-04			1423		-7794861
sim6	11091	20259	9.60E-04			1437		-7842770
sim7	11557	19589	7.20E-04			1439		-7863967
sim8	13284	19679	3.30E-03			1460		-7798626
Mean	11185	20226	1.10E-03			1438		
SD	1261	426	9.10E-04			11		

Table S13: Inference results using the 'isolation with migration' model where gene flow stops at T_M = 6 kya, for simulated data, but with larger *m*. Demographic parameters inferred by *diCal 2.0* using the IM model with gene flow stopping time (Figure S19a), for data simulated under the same model with T_{DIV} = 20 kya and m = 0.0001. Parameters with the 'mle' prefix were estimated and the other parameters were fixed.

Mle	mle T_M	mle T_{DIV}	m	N_1	N_2	mle N_B	N_A	logLike
Truth	**6000**	**20000**	**1.00E-03**	**3000**	**3000**	**1800**	**20000**	
sim1	2328	20589				1437		-7815356
sim2	1274	19858				1458		-7840936
sim3	6750	19613				1446		-7840430
sim4	6564	19900				1426		-7730805
Mean	4229	19990				1442		
SD	2837	419				14		

Table S14: **Inference results using the second contact model, for simulated data.** Demographic parameters inferred by *diCal 2.0* using the second contact model (Figure S19b), for data simulated under the same model with T_{DIV} = 35 kya, T_M = 25 kya, m_1 = 0.001, and m_2 = 0.01. Parameters with the 'mle' prefix were estimated and the other parameters were fixed.

	mle T_M	mle T_{DIV}	mle m_1	mle m_2	N_1	N_2	mle N_B	N_A	logLike
Truth	**25000**	**35000**	**1.00E-03**	**1.00E-02**	**3000**	**3000**	**1800**	**20000**	
sim1	11838	37520	4.30E-04	7.50E-04			1380		-8913128
sim2	11613	36934	5.10E-05	7.00E-04			1412		-8988046
sim3	11602	37158	2.50E-07	8.10E-04			1379		-8960848
sim4	11636	38228	5.00E-04	7.00E-04			1349		-8959106
sim5	11612	35631	3.70E-04	6.30E-04			1417		-8928802
sim6	12508	37199	1.00E-03	8.30E-04			1372		-8973057
sim7	11631	36657	3.70E-04	7.00E-04			1393		-8919797
sim8	11623	36887	3.80E-04	5.70E-04			1363		-8877822
Mean	11758	37027	3.90E-04	7.10E-04			1383		
SD	313	741	3.00E-04	8.50E-05			23		

Table S15: Inference results using the second contact model with fixed migration rates, for simulated data. Demographic parameters inferred by *diCal 2.0* using the second contact model (Figure S19b), for data simulated under the same model with T_{DIV} = 35 kya, T_M = 25 kya, m_1 = 0.001, and m_2 = 0.01. Parameters with the 'mle' prefix were estimated and the other parameters were fixed.

	mle T_M	mle T_{DIV}	m_1	m_2	N_1	N_2	mle N_B	N_A	logLike
Truth	**25000**	**35000**	**1.00E-03**	**1.00E-02**	**3000**	**3000**	**1800**	**20000**	
sim1	23055	36475					1412		-9042996
sim2	21765	37146					1404		-9070580
sim3	22140	36428					1416		-9045602
sim4	21877	36106					1406		-8950596
Mean	22209	36539					1409		
SD	585	437					6		

Table S16: Inference results using the clean split model for data simulated under the clean split model, used for estimating the likelihood ratio distribution. Demographic parameters inferred by *diCal 2.0* using the clean split model (Figure S19a, $m = 0$, no T_M) for data simulated under the same model with $T_{DIV} = 20$ kya. Parameters with the 'mle' prefix were estimated and the other parameters were fixed.

	mle T_{DIV}	N_1	N_2	mle N_B	N_A	logLike
Truth	**20000**	**3000**	**3000**	**1800**	**20000**	
sim1	20012			1414		-7837516
sim2	20022			1417		-7810549
sim3	19432			1440		-7846074
sim4	19499			1457		-7801399
sim5	19155			1459		-7872743
sim6	19029			1455		-7864050
sim7	19656			1443		-7856902
sim8	20004			1442		-7804585
sim9	19626			1444		-7778001
sim10	19824			1435		-7859782
sim11	20070			1426		-7782285
sim12	19561			1422		-7749106
sim13	19158			1449		-7803120
sim14	19440			1425		-7824226
sim15	19668			1425		-7779055
sim16	19485			1457		-7772957
sim17	20031			1439		-7868893
sim18	19424			1424		-7867072
sim19	20172			1438		-7818698
sim20	19853			1465		-7870977
Mean	19656			1439		
SD	335			15		

Table S17: Inference results using the 'isolation with migration' model for data simulated under the clean split model, used for estimating the likelihood ratio distribution. Demographic parameters inferred by *diCal 2.0* using the 'isolation with migration' model (Figure S19a, $T_M = 0$) for data simulated under the clean split model (Figure 19a, $T_M = 0$) with $T_{DIV} = 20$ kya. Parameters with the 'mle' prefix were estimated and the other parameters were fixed.

	mle T_{DIV}	mle m	N_1	N_2	mle N_B	N_A	logLike
Truth	**20000**	**0.00E+00**	**3000**	**3000**	**1800**	**20000**	
sim1	20431	1.10E-04			1410		-7837511
sim2	20149	4.50E-05			1416		-7810548
sim3	19769	1.10E-04			1438		-7846071
sim4	19834	1.10E-04			1454		-7801396
sim5	19777	2.00E-04			1453		-7872732
sim6	19097	3.10E-05			1456		-7864050
sim7	19744	2.30E-05			1443		-7856902
sim8	20236	6.80E-05			1439		-7804583
sim9	20205	1.90E-04			1439		-7777990
sim10	19914	1.40E-05			1433		-7859782
sim11	20583	1.50E-04			1421		-7782277
sim12	19612	3.80E-06			1420		-7749106
sim13	19548	1.30E-04			1446		-7803117
sim14	19485	1.10E-05			1425		-7824225
sim15	19863	5.00E-05			1423		-7779054
sim16	19730	7.90E-05			1455		-7772952
sim17	20400	1.10E-04			1436		-7868890
sim18	19618	5.90E-05			1423		-7867070
sim19	20989	2.20E-04			1430		-7818681
sim20	20372	1.60E-04			1460		-7870967
Mean	19968	9.30E-05			1436		
SD	446	6.50E-05			15		

Table S18: Inference results using the 'isolation with migration' model for simulated data. Demographic parameters inferred by *diCal 2.0* using the IM model (Figure S19a, $m = 0$, no T_M) for data simulated under $T_{DIV} = 20$ kya and $m = 0.0005$. Parameters with the 'mle' prefix were estimated and the other parameters were fixed.

	mle T_{DIV}	N_1	N_2	mle N_B	N_A	logLike
Truth	**20000**	**3000**	**3000**	**1800**	**20000**	
sim1	18391			1442		-7746001
sim2	18416			1465		-7824461
sim3	18289			1471		-7877600
sim4	16526			1471		-7779478
Mean	17905			1462		
SD	922			14		

Table S19: Inference results using the second contact model for data simulated under the 'isolation with migration' model. Demographic parameters inferred by *diCal 2.0* using the second contact model (Figure S19b) for data simulated under the IM model (FigureS19a, $T_M = 0$) with $T_{DIV} = 20$ kya and $m = 0.0005$. Parameters with the 'mle' prefix were estimated and the other parameters were fixed.

	mle T_{DIV}	mle m	N_1	N_2	mle N_B	N_A	logLike
Truth	**20000**	**5.00E-04**	**3000**	**3000**	**1800**	**20000**	
sim1	20152	8.90E-04			1426		-7745921
sim2	19918	6.50E-04			1451		-7824403
sim3	20256	1.20E-03			1454		-7877504
sim4	18996	6.80E-04			1439		-7779426
Mean	19830	8.40E-04			1443		
SD	574	2.40E-04			13		

Table S20: Log-likelihood differences between the clean split and the 'isolation with migration' model. For each pair, the log-likelihood at the estimated maximum from the analysis of the real data under a IM model (Table S5) was subtracted from that obtained using the clean split model (Table S4).

Population 1	Population 2	logLike diff
Athabascan	USR1	35.7
Aymara	USR1	33.1
Karitiana	USR1	145.7
Koryak	USR1	466.9
Nivkh	USR1	394.5
Han	USR1	181.7

Table S21: **95% confidence intervals for T_{DIV} and T_M.** We obtained confidence intervals following the parametric bootstrap approach detailed in Section 18.5. We show bias-corrected estimates for T_{DIV} and T_M, based on our simulation study Section 18.3. Note that T_M estimates should be taken with caution as these depend on a fixed migration rate. All values are shown in units of 'thousands of years ago' (kya).

Pop 1	Pop 2	T_{DIV}	T_{DIV} +/- C.I.	T_M	T_M +/- C.I
Athabascan	USR1	19.7	0.8	4.2	5.6
Aymara	USR1	21.2	0.8	13.2	5.6
Karitiana	USR1	22.6	0.8	12.1	5.6
Koryak	USR1	34.2	1.5	27.2	1.1
Nivkh	USR1	34.3	1.5	28.9	1.1
Han	USR1	33.5	1.5	29.3	1.1

Table S22: The overall site frequency of the genome sample used for inferring the backbone demography as well as the most likely join-on point for USR1. The most common types of mutations are singletons in Han and Koryaks, followed by Athabascans and then Karitiana. Athabascans have almost 50% more private mutations than Karitiana, even though they diverged most recently. Karitiana and USR1 have roughly the same number of private mutations.

Ancestral/derived					
Han	**Koryak**	**Athabascan**	**Karitiana**	**USR1**	**Kilobases**
2/0	3/1	4/0	4/0	2/0	192.31
1/1	4/0	4/0	4/0	2/0	180.66
2/0	4/0	3/1	4/0	2/0	158.73
2/0	4/0	4/0	3/1	2/0	106.97
2/0	4/0	4/0	4/0	1/1	95.58
2/0	2/2	4/0	4/0	2/0	49.73
2/0	3/1	2/0	4/0	2/0	38.31
2/0	4/0	4/0	2/2	2/0	37.31
2/0	4/0	2/2	4/0	2/0	36.25
1/1	3/1	4/0	4/0	2/0	36.09
1/1	4/0	2/0	4/0	2/0	35.57
2/0	3/1	3/1	4/0	2/0	35.3
2/0	4/0	3/1	3/1	2/0	32.32
1/1	4/0	3/1	4/0	2/0	24.78
2/0	4/0	3/1	4/0	1/1	24.35
All other configurations					4184.79
Monomorphic					2.67E06

Table S23: Bootstrap results for split times in backbone populations. 102 datasets were simulated under the MLEs and the whole inference process described in Section 19 was repeated on each replicate. The distribution of the split times in the backbone demography for all replicates is summarized below.

	Atha-Kari	Koryak-AK	Han-Native American
count	103	103	103
mean	17435	25950.1	28174.9
std	738.113	1330.78	1094.98
min	16314.3	24021.7	25719.4
50%	17304.3	25604.5	28197.2
max	19936.1	30153.5	31153.5
95% CI	[1.64e+04, 1.93e+04]	[2.46e+04, 2.99e+04]	[2.64e+04, 3.09e+04]
Truth	15754.3	23260.3	25588.8

Table S24: **Bootstrap results for migration rates in backbone populations.** 102 datasets were simulated under the MLEs and the whole inference process described in Section 19 was repeated on each replicate. The distribution of the migration rates in the backbone demography for all replicates is summarized below.

	Koryak-AK	Han-Koryak	Atha-Kari	Koryak-Atha. (2nd contact)
count	103	103	103	103
mean	-2.42822	-1.09435	-1.78372	-2.21815
std	1.05531	0.157509	0.541112	0.953633
min	-5	-1.73998	-4.73775	-5
50%	-2.13459	-1	-1.75108	1.84707
max	-1	-1	-1	-1.16011
95% CI	[-4.96, -1]	[-1.6, -1]	[-3.18, -1]	[-5, -1.21]
Truth	-2.67006	-2.98434	-3.54912	-2.66249

Table S25: Bootstrap results for the USR1 join-on time. 102 datasets were simulated under the MLEs and the whole inference process described in Section 19 was repeated on each replicate. The distribution of the difference between the estimated join-on times and the MLE for all replicates is summarized below.

	Split(Koryak/AK) - MLE
count	103
mean	1630.55
std	917.899
min	468.255
50%	1448.77
max	5411.47
95% CI	[576, 4.39e+03]

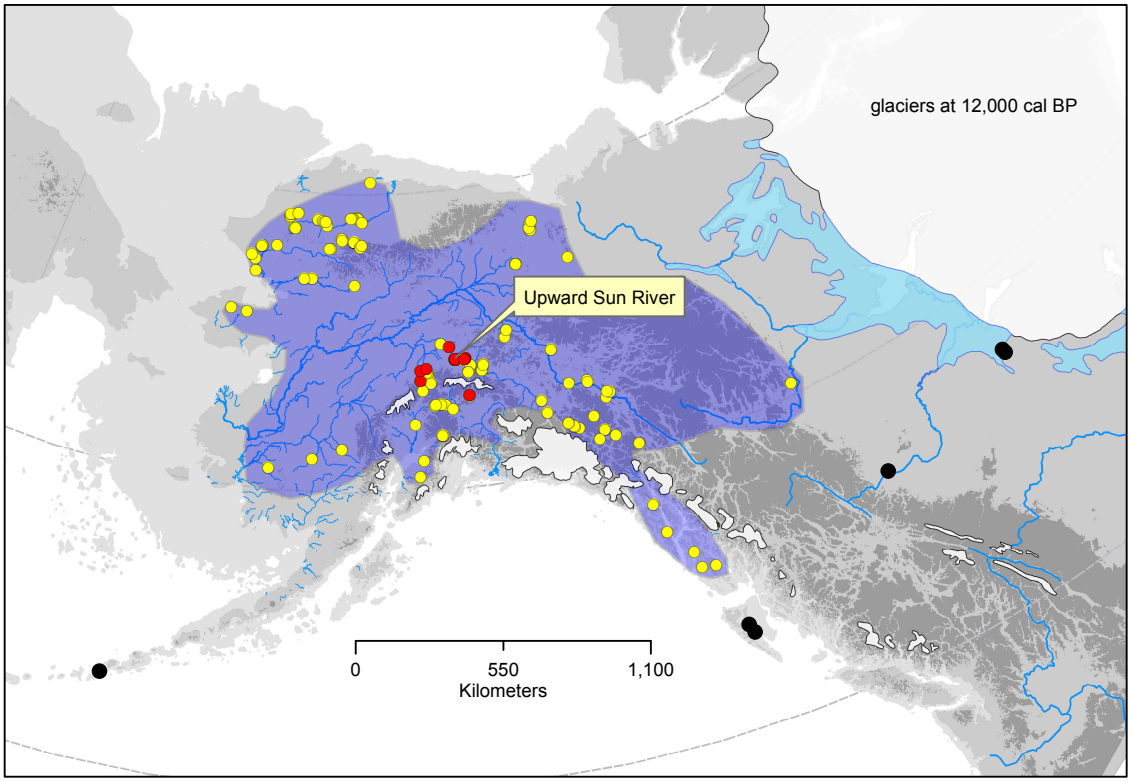
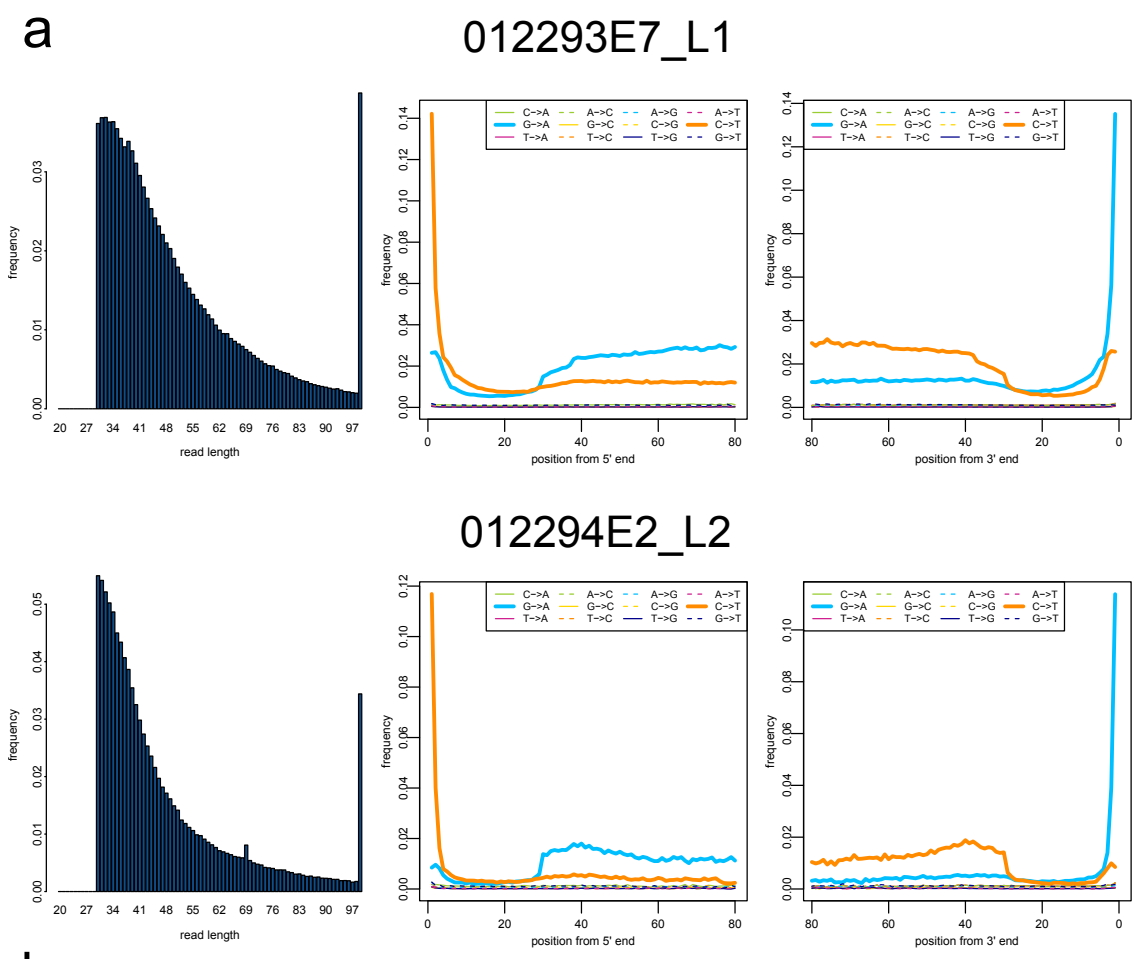

glaciers at 12,000 cal BP
Upward Sun River
0
550
1,100
Kilometers

Figure S1. Denali complex extent in Beringia. Geographic extent derived from sites attributed to the Denali complex or American Paleoarctic tradition (yellow points), and other sites with wedge-shaped microblade cores (black points). Early Denali complex sites (pre 11 kya) are illustrated as red points. Glacial limits and proglacial lakes are derived from [133].


a
012293E7_L1
frequency
read length
position from 5' end
position from 3' end
012294E2_L2
frequency
read length
position from 5' end
position from 3' end

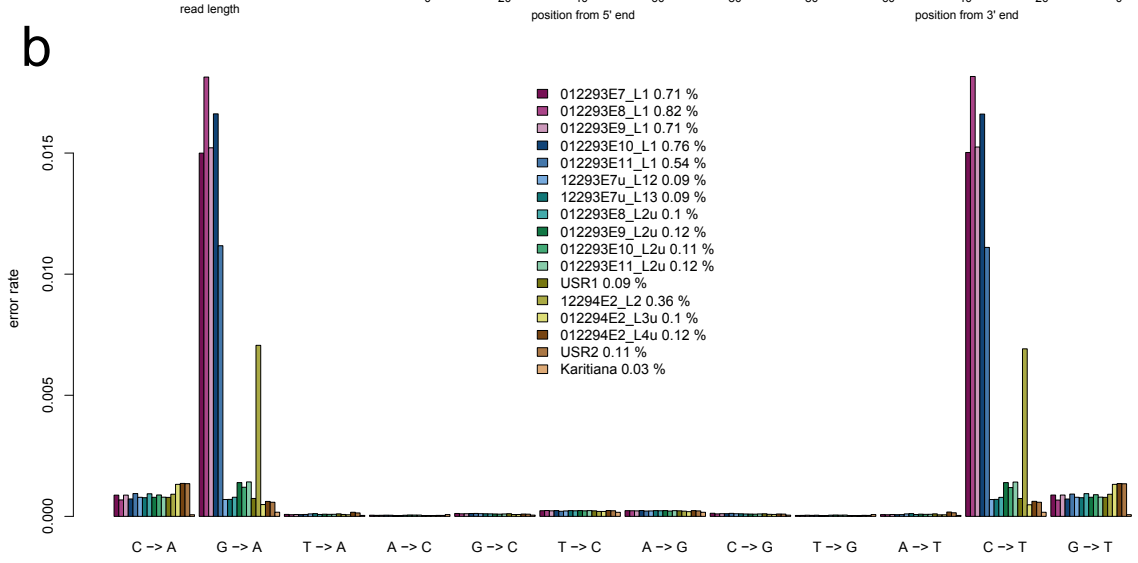

b
error rate
012293E7_L1 0.71 %
012293E8_L1 0.82 %
012293E9_L1 0.71 %
012293E10_L1 0.76 %
012293E11_L1 0.54 %
12293E7u_L12 0.09 %
12293E7u_L13 0.09 %
012293E8_L2u 0.1 %
012293E9_L2u 0.12 %
012293E10_L2u 0.11 %
012293E11_L2u 0.12 %
USR1 0.09 %
12294E2_L2 0.36 %
012294E2_L3u 0.1 %
012294E2_L4u 0.12 %
USR2 0.11 %
Karitiana 0.03 %
C -> A
G -> A
T -> A
A -> C
G -> C
T -> C
A -> G
C -> G
T -> G
A -> T
C -> T
G -> T

Figure S2. Error patterns in sequencing data. a. Left: Fragment length distribution after removing adapter sequences and collapsing paired-end reads. Right: Misincorporation frequency (y-axis) with respect to the position (x-axis) from the 5' and 3' ends of the collapsed reads for one USR1 (top) and one USR2 (bottom) non USER-treated library. **b.** Relative error rates (excess derived alleles compared to a high coverage 1000 genomes individual) in sequencing data from all USR1 and USR2 libraries, as well as a modern-day Native American genome for comparison. The overall relative error rates for each library are shown in the middle.

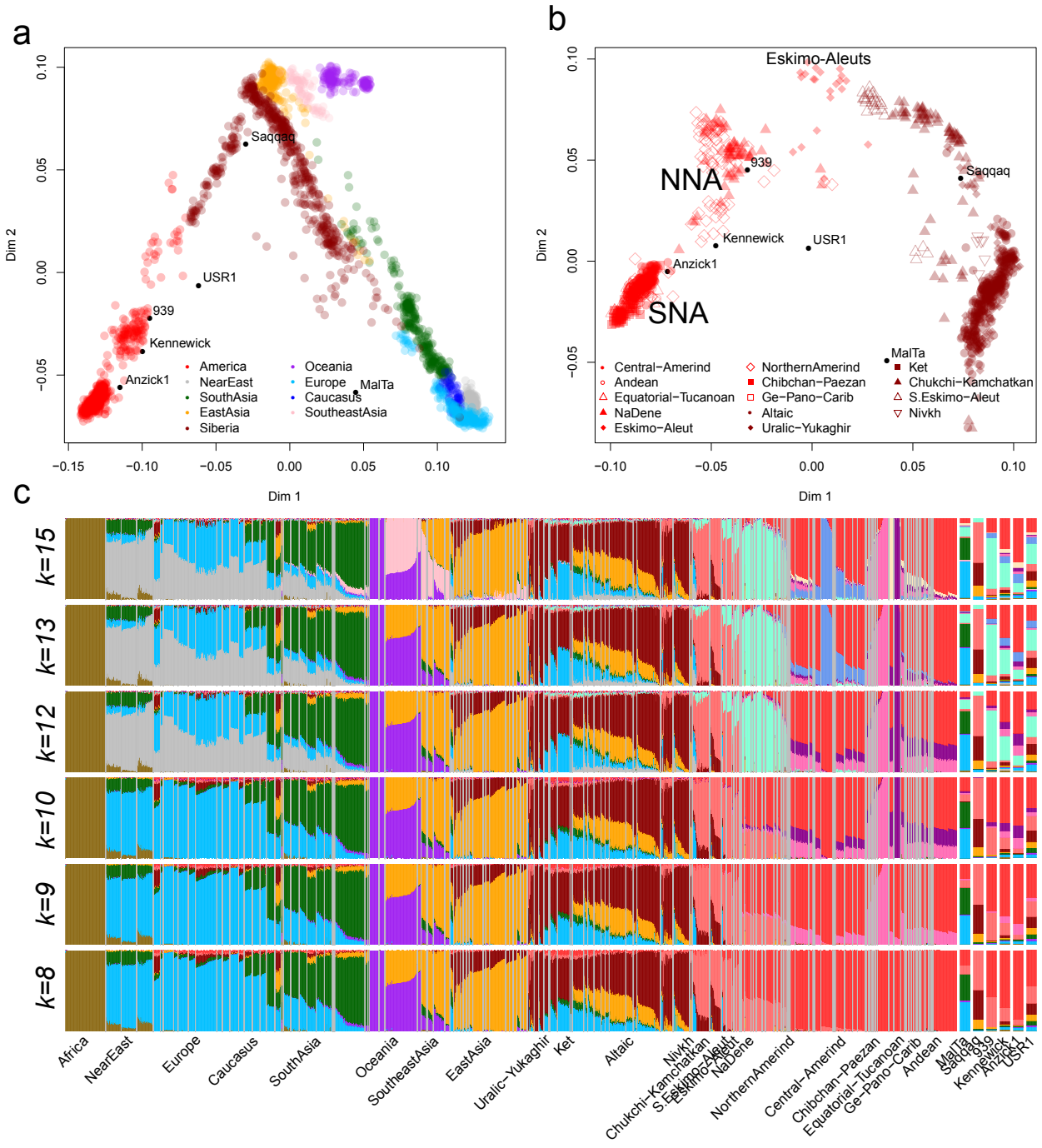

a
Dim 2
Dim 1
Saqqaq
USR1
939
Kennewick
Anzick1
MalTa
America
NearEast
SouthAsia
EastAsia
Siberia
Oceania
Europe
Caucasus
SoutheastAsia
b
Eskimo-Aleuts
NNA
SNA
939
Saqqaq
Kennewick
USR1
Anzick1
MalTa
Central−Amerind
Andean
Equatorial−Tucanoan
NaDene
Eskimo−Aleut
NorthernAmerind
Chibchan−Paezan
Ge−Pano−Carib
Altaic
Uralic−Yukaghir
Ket
Chukchi−Kamchatkan
S.Eskimo−Aleut
Nivkh
c
k=15
k=13
k=12
k=10
k=9
k=8
Africa
NearEast
Europe
Caucasus
SouthAsia
Oceania
SoutheastAsia
EastAsia
Uralic−Yukaghir
Ket
Altaic
Nivkh
Chukchi−Kamchatkan
S.Eskimo−Aleut
NaDene
NorthernAmerind
Central−Amerind
Chibchan−Paezan
Equatorial−Tucanoan
Ge−Pano−Carib
Andean
MalTa
Saqqaq
939
Kennewick
Anzick1
USR1

Figure S3. Genetic relationships between USR1 and modern-day worldwide populations. USR1 and other five ancient genomes (Anzick1, 939, Kennewick, Saqqaq and Mal'ta) were compared to a genotype panel of 2,537 modern-day individuals from 167 ethnic groups, genotyped over 199,285 SNP sites. **a.** Multidimensional scaling (MDS) plot for all non-African individuals in the genotype panel. Individuals are colored based on their broad continental origin. **b.** MDS plot for all Native American and Siberian individuals in the genotype panel. Individuals were assigned using the language-based classification described in SI Section 6.1. In both plots, the first two eigenvectors are shown in the x- and y- axis, respectively. **c.** *ADMIXTURE* analysis assuming *K={8,9,10,12,13,15}* ancestral components. Admixture proportions for the six ancient genomes were estimated based on the allele frequency estimates produced by *ADMIXTURE*. Each bar represents an individual and each color represents the admixture proportion from each of the K assumed ancestral components.

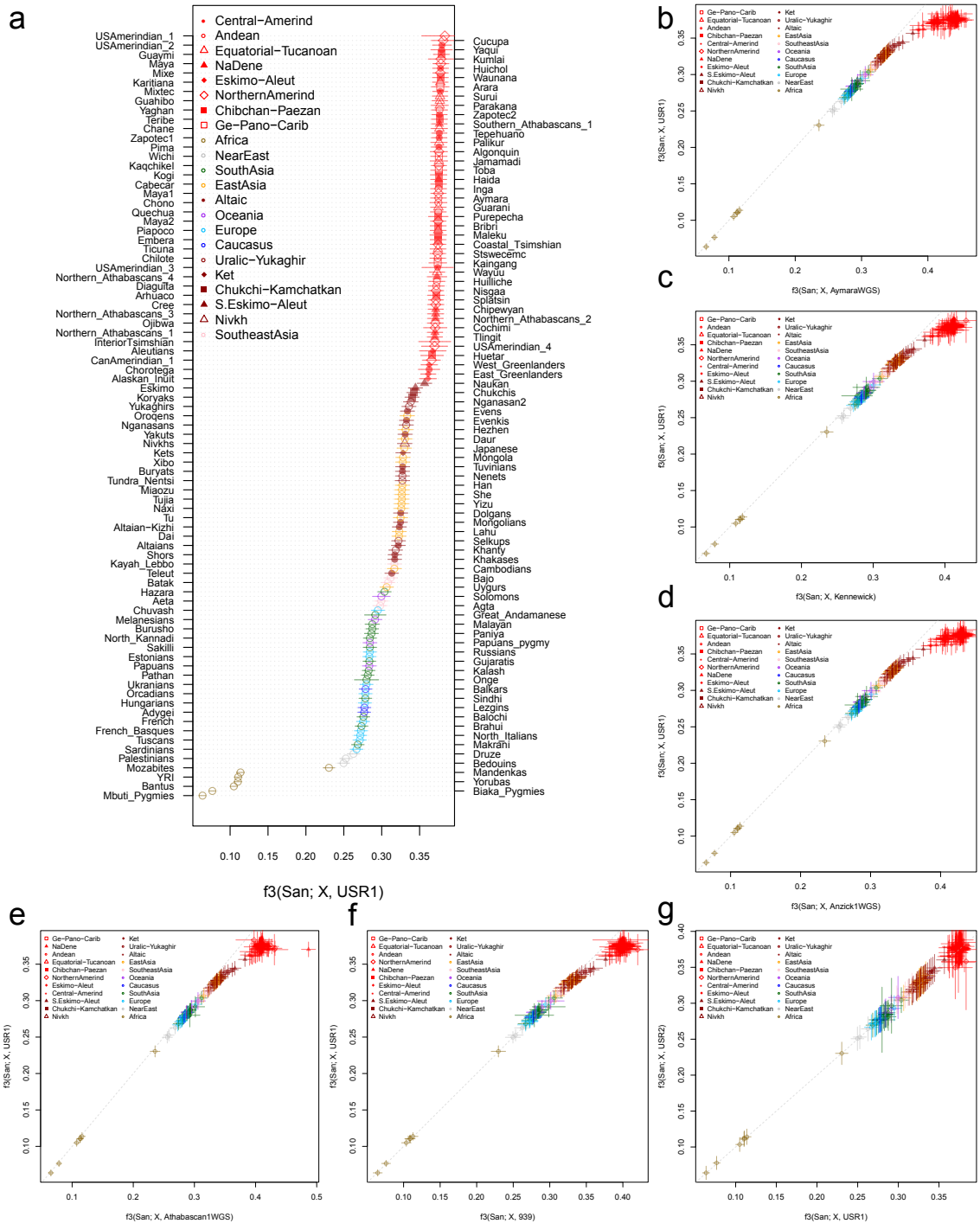

a
Central−Amerind
Andean
Equatorial−Tucanoan
NaDene
Eskimo−Aleut
NorthernAmerind
Chibchan−Paezan
Ge−Pano−Carib
Africa
NearEast
SouthAsia
EastAsia
Altaic
Oceania
Europe
Caucasus
Uralic−Yukaghir
Ket
Chukchi−Kamchatkan
S.Eskimo−Aleut
Nivkh
SoutheastAsia
USAmerindian_1
USAmerindian_2
Guaymi
Maya
Mixe
Karitiana
Mixtec
Guahibo
Yaghan
Teribe
Chane
Zapotec1
Pima
Wichi
Kaqchikel
Kogi
Cabecar
Maya1
Chono
Quechua
Maya2
Piapoco
Embera
Ticuna
Chilote
USAmerindian_3
Northern_Athabascans_4
Diaguita
Arhuaco
Cree
Northern_Athabascans_3
Ojibwa
Northern_Athabascans_1
InteriorTsimshian
Aleutians
CanAmerindian_1
Chorotega
Alaskan_Inuit
Eskimo
Koryaks
Yukaghirs
Orogens
Nganasans
Yakuts
Nivkhs
Kets
Xibo
Buryats
Tundra_Nentsi
Miaozu
Tujia
Naxi
Tu
Altaian−Kizhi
Dai
Altaians
Shors
Kayah_Lebbo
Teleut
Batak
Hazara
Aeta
Chuvash
Melanesians
Burusho
North_Kannadi
Sakilli
Estonians
Papuans
Pathan
Ukranians
Orcadians
Hungarians
Adygei
Frerich
French_Basques
Tuscans
Sardinians
Palestinians
Mozabites
YRI
Bantus
Mbuti_Pygmies
Cucupa
Yaqui
Kumlai
Huichol
Waunana
Arara
Surui
Parakana
Zapotec2
Southern_Athabascans_1
Tepehuano
Palikur
Algonquin
Jamamadi
Toba
Haida
Inga
Aymara
Guarani
Purepecha
Bribri
Maleku
Coastal_Tsimshian
Stswecēmc
Kaingang
Wayuu
Huilliche
Nisgaa
Splatsin
Chipewyan
Northern_Athabascans_2
Cochimi
Tlingit
USAmerindian_4
Huetar
West_Greenlanders
East_Greenlanders
Naukan
Chukchis
Nganasan2
Evens
Evenkis
Hezhen
Daur
Japanese
Mongola
Tuvinians
Nenets
Han
She
Yizu
Dolgans
Mongolians
Lahu
Selkups
Khanty
Khakases
Cambodians
Bajo
Uygurs
Solomons
Agta
Great_Andamanese
Malayan
Paniya
Papuans_pygmy
Russians
Gujaratis
Kalash
Onge
Balkars
Sindhi
Lezgins
Balochi
Brahui
North_Italians
Makrani
Druze
Bedouins
Mandenkas
Yorubas
Biaka_Pygmies
0.10 0.15 0.20 0.25 0.30 0.35
f3(San; X, USR1)
b
f3(San; X, USR1)
0.1 0.2 0.3 0.4
f3(San; X, AymaraWGS)
c
f3(San; X, USR1)
0.1 0.2 0.3 0.4
f3(San; X, Kennewick)
d
f3(San; X, USR1)
0.1 0.2 0.3 0.4
f3(San; X, Anzick1WGS)
e
f3(San; X, USR1)
0.1 0.2 0.3 0.4 0.5
f3(San; X, Athabascan1WGS)
f
f3(San; X, USR1)
0.10 0.15 0.20 0.25 0.30 0.35 0.40
f3(San; X, 939)
g
f3(San; X, USR2)
0.10 0.15 0.20 0.25 0.30 0.35
f3(San; X, USR1)

Figure S4. Shared genetic drift between USR1 and modern-day worldwide populations. USR1 and other modern and ancient genomes (Aymara, Athabascan1, Anzick1, 939, Kennewick and USR2) were compared to a genotype panel including a worldwide sample of 167 ethnic groups, genotyped over 199,285 SNP sites, using outgroup f3 statistics of the form *f3(San; whole-genome, X)*. **a.** Results for *f3(San; USR1, X)*. Error bars show one standard error. USR1 is most closely related to modern-day Native American populations. **b-g.** We compared the *f3* statistics vector obtained for USR1 to those obtained for the other whole genomes. In each panel, statistics of the form *f3(San; whole-genome, X)* are shown in the x-axis and *f3(San; USR1, X)* is shown on the y-axis. Old-World populations are colored based on their broad geographic origin and Native American populations were assigned using the language-based classification described in SI Section 6.1. Points show *f3* statistics, and error bars show ~2.57 standard errors, which corresponds to a p-value of 0.01 in a Z-test.

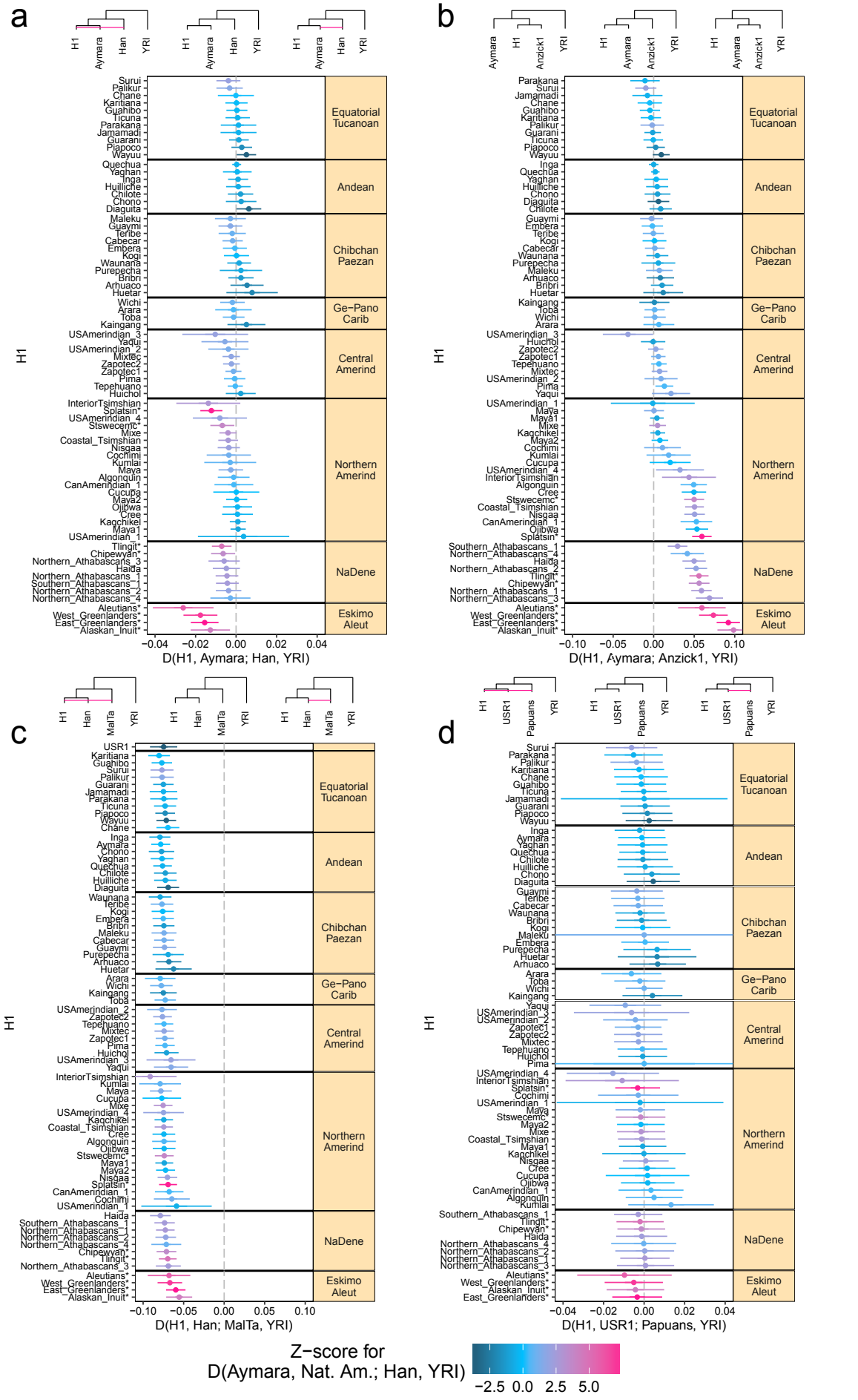

a
H1
Aymara
Han
YRI
Surui
Palikur
Chane
Karitiana
Guahibo
Ticuna
Parakana
Jamamadi
Guarani
Piapoco
Wayuu
Quechua
Yaghan
Inga
Huilliche
Chilote
Chono
Diaguita
Maleku
Guaymi
Teribe
Cabecar
Embera
Kogi
Waunana
Purepecha
Bribri
Arhuaco
Huetar
Wichi
Arara
Toba
Kaingang
USAmerindian_3
Yaqui
USAmerindian_2
Mixtec
Zapotec2
Zapotec1
Pima
Tepehuano
Huichol
InteriorTsimshian
Splatsin*
USAmerindian_4
Stswecemc*
Mixe
Coastal_Tsimshian
Nisgaa
Cochimi
Kumlai
Maya
Algonquin
CanAmerindian_1
Cucupa
Maya2
Ojibwa
Cree
Kaqchikel
Maya1
USAmerindian_1
Tlingit*
Chipewyan*
Northern_Athabascans_3
Haida
Northern_Athabascans_1
Southern_Athabascans_1
Northern_Athabascans_2
Northern_Athabascans_4
Aleutians*
West_Greenlanders*
East_Greenlanders*
Alaskan_Inuit*
Equatorial Tucanoan
Andean
Chibchan Paezan
Ge−Pano Carib
Central Amerind
Northern Amerind
NaDene
Eskimo Aleut
−0.04 −0.02 0.00 0.02 0.04
D(H1, Aymara; Han, YRI)
b
Aymara
H1
Anzick1
YRI
Parakana
Surui
Jamamadi
Chane
Guahibo
Karitiana
Palikur
Guarani
Ticuna
Piapoco
Wayuu
Inga
Quechua
Yaghan
Huilliche
Chono
Diaguita
Chilote
Guaymi
Embera
Teribe
Kogi
Cabecar
Waunana
Purepecha
Maleku
Arhuaco
Bribri
Huetar
Kaingang
Toba
Wichi
Arara
USAmerindian_3
Huichol
Zapotec2
Zapotec1
Tepehuano
Mixtec
USAmerindian_2
Pima
Yaqui
USAmerindian_1
Maya
Maya1
Mixe
Kaqchikel
Maya2
Cochimi
Kumlai
Cucupa
USAmerindian_4
InteriorTsimshian
Algonquin
Cree
Stswecemc*
Coastal_Tsimshian
Nisgaa
CanAmerindian_1
Ojibwa
Splatsin*
Southern_Athabascans_1
Northern_Athabascans_4
Haida
Northern_Athabascans_2
Tlingit*
Chipewyan*
Northern_Athabascans_1
Northern_Athabascans_3
Aleutians*
West_Greenlanders*
East_Greenlanders*
Alaskan_Inuit*
Equatorial Tucanoan
Andean
Chibchan Paezan
Ge−Pano Carib
Central Amerind
Northern Amerind
NaDene
Eskimo Aleut
−0.10 −0.05 0.00 0.05 0.10
D(H1, Aymara; Anzick1, YRI)
c
H1
Han
MalTa
YRI
USR1
Karitiana
Guahibo
Surui
Palikur
Guarani
Jamamadi
Parakana
Ticuna
Piapoco
Wayuu
Chane
Inga
Aymara
Chono
Yaghan
Quechua
Chilote
Huilliche
Diaguita
Waunana
Teribe
Kogi
Embera
Bribri
Maleku
Cabecar
Guaymi
Purepecha
Arhuaco
Huetar
Arara
Wichi
Kaingang
Toba
USAmerindian_2
Zapotec2
Tepehuano
Mixtec
Zapotec1
Pima
Huichol
USAmerindian_3
Yaqui
InteriorTsimshian
Kumlai
Maya
Cucupa
Mixe
USAmerindian_4
Kaqchikel
Coastal_Tsimshian
Cree
Algonquin
Ojibwa
Stswecemc*
Maya1
Maya2
Nisgaa
Splatsin*
CanAmerindian_1
Cochimi
USAmerindian_1
Haida
Southern_Athabascans_1
Northern_Athabascans_1
Northern_Athabascans_2
Northern_Athabascans_4
Chipewyan*
Tlingit*
Northern_Athabascans_3
Aleutians*
West_Greenlanders*
East_Greenlanders*
Alaskan_Inuit*
Equatorial Tucanoan
Andean
Chibchan Paezan
Ge−Pano Carib
Central Amerind
Northern Amerind
NaDene
Eskimo Aleut
−0.10 −0.05 0.00 0.05 0.10
D(H1, Han; MalTa, YRI)
d
H1
USR1
Papuans
YRI
Surui
Parakana
Palikur
Karitiana
Chane
Guahibo
Ticuna
Jamamadi
Guarani
Piapoco
Wayuu
Inga
Aymara
Yaghan
Quechua
Chilote
Huilliche
Chono
Diaguita
Guaymi
Teribe
Cabecar
Waunana
Bribri
Kogi
Maleku
Embera
Purepecha
Huetar
Arhuaco
Arara
Toba
Wichi
Kaingang
Yaqui
USAmerindian_3
USAmerindian_2
Zapotec1
Zapotec2
Mixtec
Tepehuano
Huichol
Pima
USAmerindian_4
InteriorTsimshian
Splatsin*
Cochimi
USAmerindian_1
Maya
Stswecemc*
Maya2
Mixe
Coastal_Tsimshian
Maya1
Kaqchikel
Nisgaa
Cree
Cucupa
Ojibwa
CanAmerindian_1
Algonquin
Kumlai
Southern_Athabascans_1
Tlingit*
Chipewyan*
Haida
Northern_Athabascans_4
Northern_Athabascans_2
Northern_Athabascans_1
Northern_Athabascans_3
Aleutians*
West_Greenlanders*
Alaskan_Inuit*
East_Greenlanders*
Equatorial Tucanoan
Andean
Chibchan Paezan
Ge−Pano Carib
Central Amerind
Northern Amerind
NaDene
Eskimo Aleut
−0.04 −0.02 0.00 0.02 0.04
D(H1, USR1; Papuans, YRI)
Z−score for
D(Aymara, Nat. Am.; Han, YRI)
−2.5 0.0 2.5 5.0

Figure S5. Allele frequency-based *D* statistics. a. We tested which Native American populations in the genotype panel carried Asian gene flow by computing *D(H1, Aymara; Han, YRI)*. In all panels, Native American populations with a statistically significant deviation from D=0 in this test are identified with a * in the y-axis. **b.** We assigned Native American populations to the North (NNA) and South (SNA) Native American lineage by computing *D(H1, Aymara; Anzick1, YRI)*. Populations with statistical support for D>0 were assigned to the NNA lineage. **c.** We assessed whether USR1 carried Ancient North Eurasian admixture as it has been shown in other Native American populations by computing *D(H1, Han; MalTa, YRI)*. USR1 and other Native American populations yielded similar results for this test. **d.** We explored if USR1 and other Native American populations were equally related to Australasians by computing D(H1, USR1; Papuans, YRI). We found no support for USR1 being closer to Papuans than other Native Americans. For all tests, we used the genotype panel described in Section 6.1. In all plots, a schematic representation of the null hypothesis for the Z-test is shown in the middle (*D=0*) and the two possible outcomes of the alternative hypothesis are shown to the left (D<0) and the right (D>0). Points represent *D* statistics and error bars represent one and ~*3.3* standard errors (which corresponds to a *p*-value of ~0.001). Native American populations were assigned using the language-based classification described in SI Section 6.1.

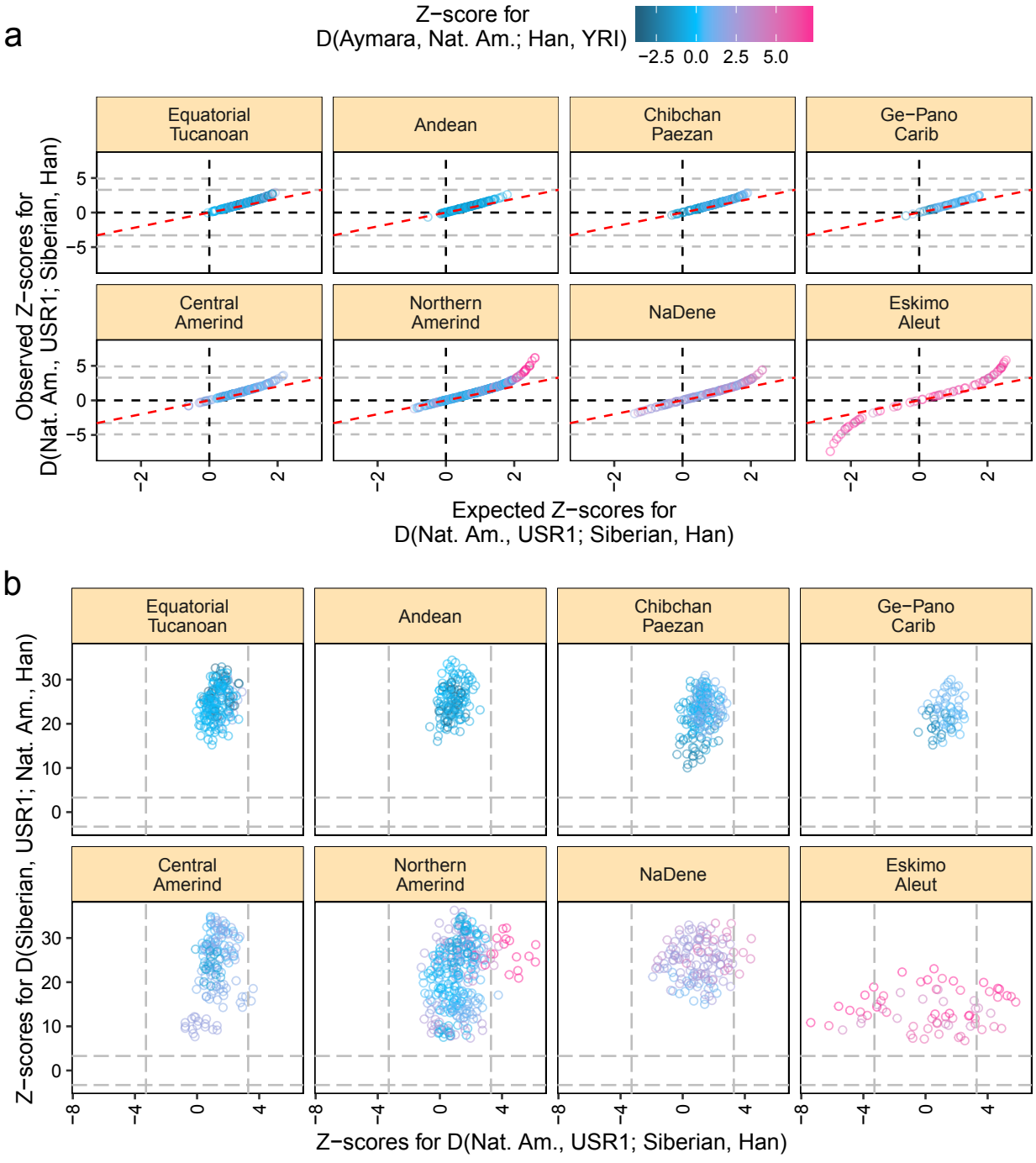

a
Z−score for
D(Aymara, Nat. Am.; Han, YRI)
−2.5 0.0 2.5 5.0
Equatorial Tucanoan
Andean
Chibchan Paezan
Ge−Pano Carib
Central Amerind
Northern Amerind
NaDene
Eskimo Aleut
Observed Z−scores for
D(Nat. Am., USR1; Siberian, Han)
Expected Z−scores for
D(Nat. Am., USR1; Siberian, Han)
b
Z−scores for D(Siberian, USR1; Nat. Am., Han)
Z−scores for D(Nat. Am., USR1; Siberian, Han)

Figure S6. Allele frequency-based D statistics testing for the relationship between USR1, Native Americans and Siberians. We tested if USR1 and Native Americans in the genotype panel formed a clade to the exclusion of other Siberian and Asian populations by computing all 11,322 possible tests of the form *D(Native American, USR1; Siberian/Han, Siberian/Han).* **a.** Quantile-quantile plot comparing the observed Z-scores to the expected normal distribution. Note that the expected distribution was obtained for the whole set of 11,322 tests, but we only show results for the tests include the Han Chinese as one of the outgroups; tests of the form *D(Native American, USR1; Siberian, Siberian)* are shown in Figure 1c. **b.** We formally tested if USR1 formed a clade with any Siberian population in the genotype panel by computing *D(Siberian, USR1; Nat. Am., Han)*. We compared these results to the above tests (*D(USR1, Native American; Siberian, Han)*). In both panels, Native American populations are colored according to the Z-score of the test *D(H1, Aymara; Han, YRI)* (SI Section 10.1), For all tests, we used the genotype panel described in Section 6.1 and populations were assigned using the language-based classification described therein. Thick dotted lines represent a *Z*-score of ~*3.3*, which corresponds to a *p*-value of ~0.001. Thin dotted lines represent a *Z*-score of ~4.91, which corresponds to a *p*-value of 0.01 after applying a Bonferroni correction for 11,322 tests.

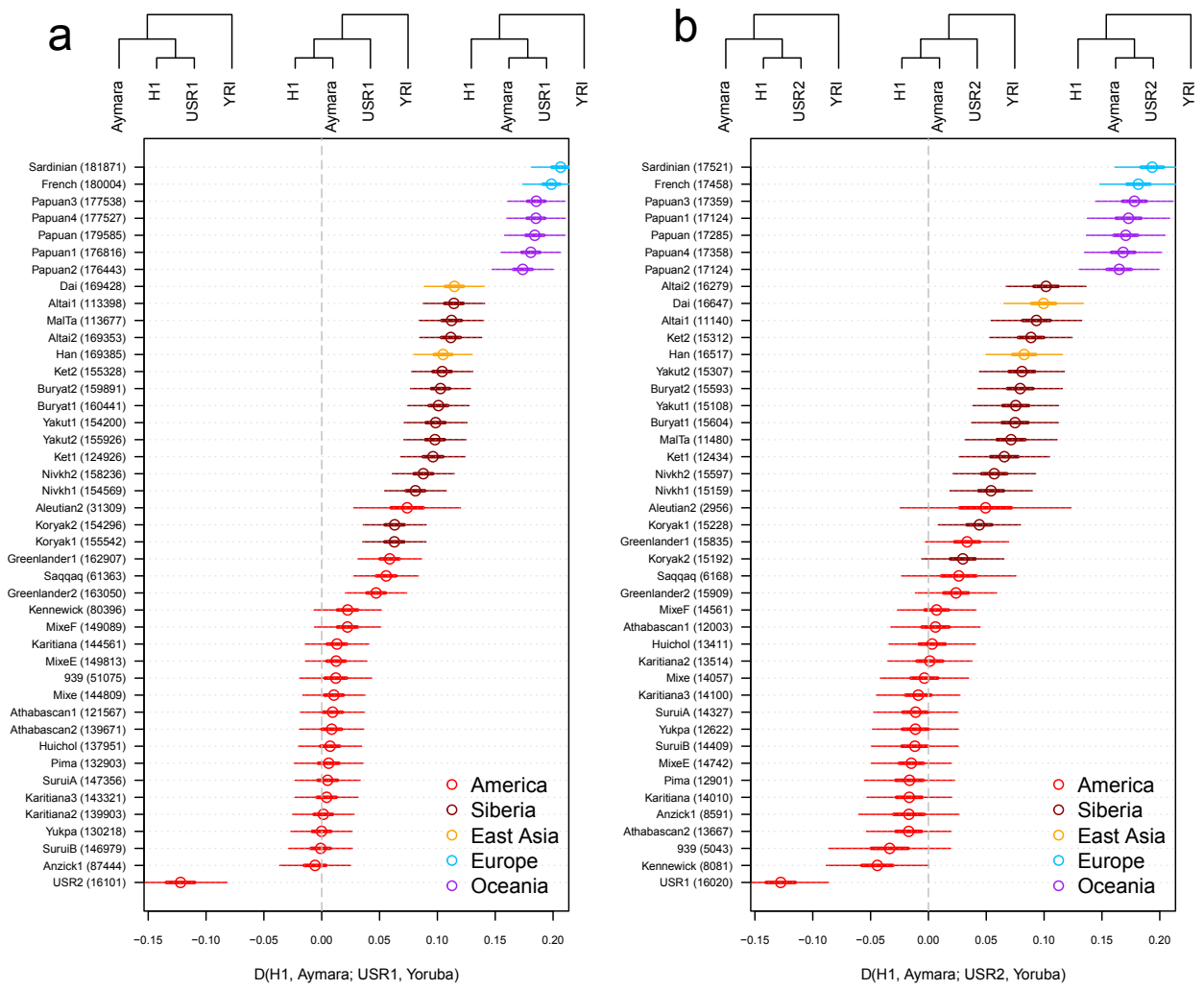

a
Aymara
H1
USR1
YRI
H1
Aymara
USR1
YRI
H1
Aymara
USR1
YRI
Sardinian (181871)
French (180004)
Papuan3 (177538)
Papuan4 (177527)
Papuan (179585)
Papuan1 (176816)
Papuan2 (176443)
Dai (169428)
Altai1 (113398)
MalTa (113677)
Altai2 (169353)
Han (169385)
Ket2 (155328)
Buryat2 (159891)
Buryat1 (160441)
Yakut1 (154200)
Yakut2 (155926)
Ket1 (124926)
Nivkh2 (158236)
Nivkh1 (154569)
Aleutian2 (31309)
Koryak2 (154296)
Koryak1 (155542)
Greenlander1 (162907)
Saqqaq (61363)
Greenlander2 (163050)
Kennewick (80396)
MixeF (149089)
Karitiana (144561)
MixeE (149813)
939 (51075)
Mixe (144809)
Athabascan1 (121567)
Athabascan2 (139671)
Huichol (137951)
Pima (132903)
SuruiA (147356)
Karitiana3 (143321)
Karitiana2 (139903)
Yukpa (130218)
SuruiB (146979)
Anzick1 (87444)
USR2 (16101)
America
Siberia
East Asia
Europe
Oceania
−0.15
−0.10
−0.05
0.00
0.05
0.10
0.15
0.20
D(H1, Aymara; USR1, Yoruba)
b
Aymara
H1
USR2
YRI
H1
Aymara
USR2
YRI
H1
Aymara
USR2
YRI
Sardinian (17521)
French (17458)
Papuan3 (17359)
Papuan1 (17124)
Papuan (17285)
Papuan4 (17358)
Papuan2 (17124)
Altai2 (16279)
Dai (16647)
Altai1 (11140)
Ket2 (15312)
Han (16517)
Yakut2 (15307)
Buryat2 (15593)
Yakut1 (15108)
Buryat1 (15604)
MalTa (11480)
Ket1 (12434)
Nivkh2 (15597)
Nivkh1 (15159)
Aleutian2 (2956)
Koryak1 (15228)
Greenlander1 (15835)
Koryak2 (15192)
Saqqaq (6168)
Greenlander2 (15909)
MixeF (14561)
Athabascan1 (12003)
Huichol (13411)
Karitiana2 (13514)
Mixe (14057)
Karitiana3 (14100)
SuruiA (14327)
Yukpa (12622)
SuruiB (14409)
MixeE (14742)
Pima (12901)
Karitiana (14010)
Anzick1 (8591)
Athabascan2 (13667)
939 (5043)
Kennewick (8081)
USR1 (16020)
America
Siberia
East Asia
Europe
Oceania
−0.15
−0.10
−0.05
0.00
0.05
0.10
0.15
0.20
D(H1, Aymara; USR2, Yoruba)

Figure S7. Whole genome-based *D* statistics. We computed *D*-statistics of the form *D(H1, Aymara; USRx, Yoruba)* to learn about the genetic relationship between the USR individuals and other modern and ancient whole genomes. For each test, one allele from each one of the four genomes was randomly sampled and transition SNPs were excluded; the total number of 'ABBA+BABA' sites is shown in brackets. A schematic representation of the null hypothesis for the Z-test is shown in the middle (*D=0*) and the two possible outcomes of the alternative hypothesis are shown to the left (D<0) and the right (D>0). Points represent *D*-statistics and error bars represent one and ~*3.3* standard errors (which corresponds to a *p*-value of ~0.001).

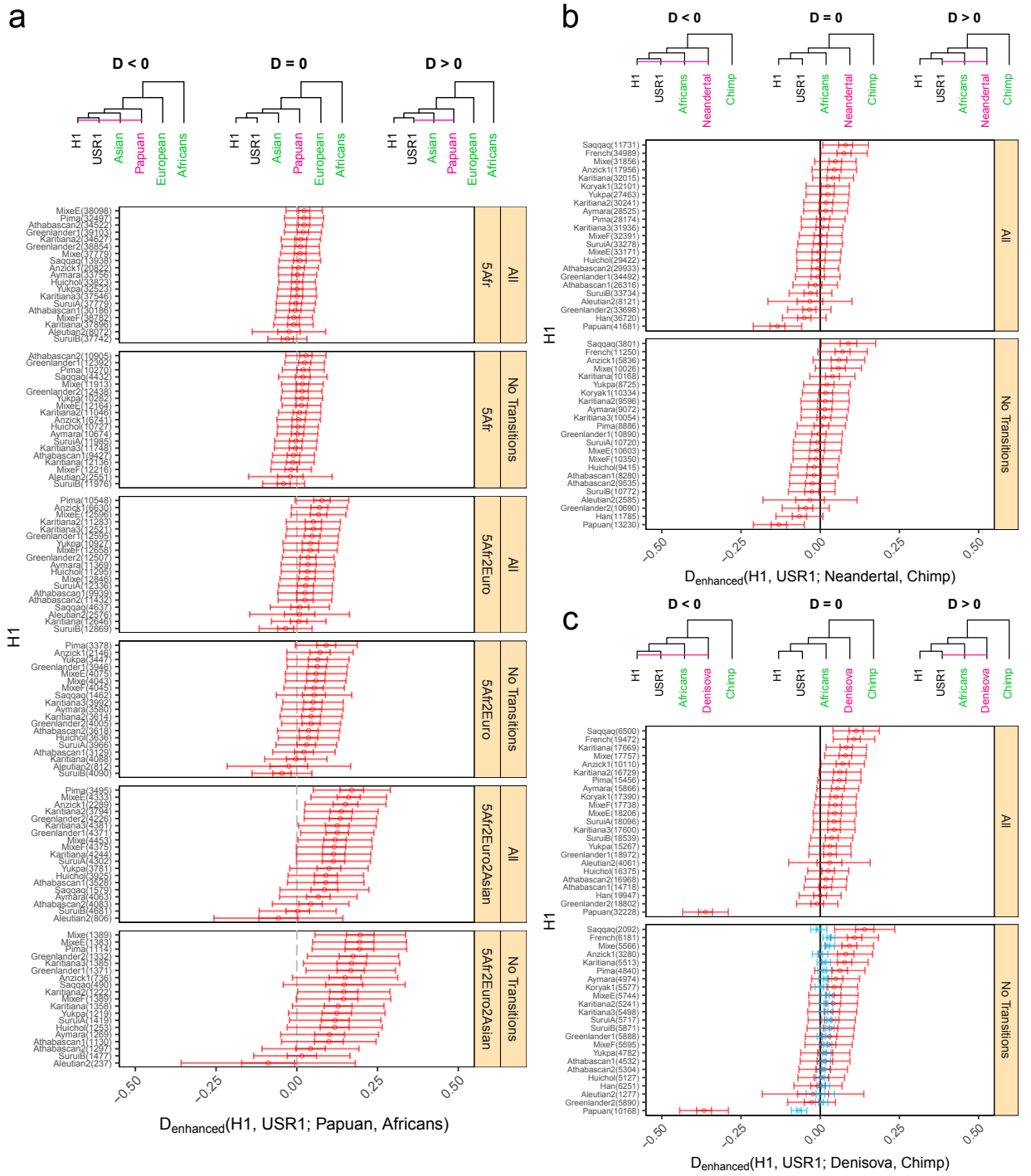

a
D < 0
D = 0
D > 0
H1
USR1
Asian
Papuan
European
Africans
5Afr
5Afr2Euro
5Afr2Euro2Asian
All
No Transitions
H1
−0.50
−0.25
0.00
0.25
0.50
Denhanced(H1, USR1; Papuan, Africans)
b
D < 0
D = 0
D > 0
H1
USR1
Africans
Neandertal
Chimp
All
No Transitions
H1
−0.50
−0.25
0.00
0.25
0.50
Denhanced(H1, USR1; Neandertal, Chimp)
c
D < 0
D = 0
D > 0
H1
USR1
Africans
Denisova
Chimp
All
No Transitions
H1
−0.50
−0.25
0.00
0.25
0.50
Denhanced(H1, USR1; Denisova, Chimp)

Figure S8. Enhanced *D* statistics show excess Denisovan ancestry in USR1. We computed 'enhanced' *D* statistics on single genomes based on different ascertainment schemes (SI Section 11) to test whether USR1 carried excess Australasian and/or archaic ancestry with respect to other Native Americans. **a.** Test for Australasian admixture in USR1, using a Papuan individual as a proxy for Australasians. **b.** Test for Neanderthal admixture in USR1, using the high-coverage Altai Neanderthal as a proxy for Neanderthals. **c.** Test for Denisovan admixture. For this test, we also show in blue the results for the 'basic' D statistics (no ascertainment step in the five African genomes). For all tests, one allele from each of the two ingroups was randomly sampled and each test was run including and excluding transition SNPs; the total number of 'ABBA+BABA' sites is shown in brackets. Note that the 'outgroup' genomes (specified in the lateral panels) represent populations that are symmetrical (an outgroup) to the *(H1, H2)* clade and that said genomes do not necessarily form a clade to the exclusion of *H3*. Rather, they are grouped to indicate that the test is restricted to sites where they carry an allele different from that in H3 (SI Section 11.1). A schematic representation (unrooted tree) of the null hypothesis for the Z-test is shown in the middle (*D=0*) and the two possible outcomes of the alternative hypothesis are shown to the left (D<0) and the right (D>0). These schematics also show the ascertainment step for each test, where sites are chosen such that the green lineages carry an allele different to that in the pink lineage. Points represent *D* statistics and error bars represent one and *~3.3* standard errors (which corresponds to a *p*-value of ~0.001).

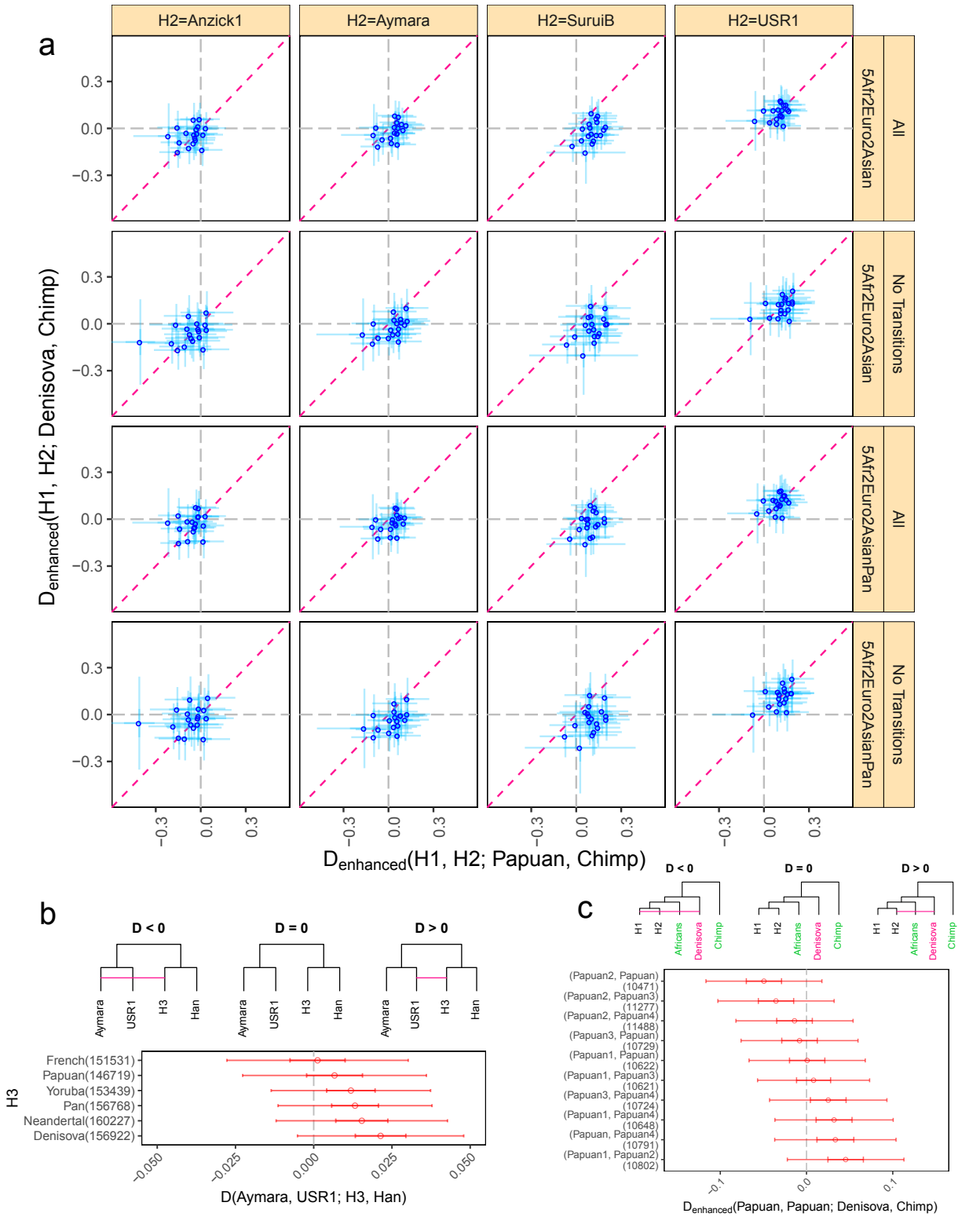

a
H2=Anzick1
H2=Aymara
H2=SuruiB
H2=USR1
5Afr2Euro2Asian
All
No Transitions
5Afr2Euro2AsianPan
Denhanced(H1, H2; Denisova, Chimp)
Denhanced(H1, H2; Papuan, Chimp)
b
D < 0
D = 0
D > 0
Aymara
USR1
H3
Han
H3
French(151531)
Papuan(146719)
Yoruba(153439)
Pan(156768)
Neandertal(160227)
Denisova(156922)
D(Aymara, USR1; H3, Han)
c
H1
H2
Africans
Denisova
Chimp
(Papuan2, Papuan) (10471)
(Papuan2, Papuan3) (11277)
(Papuan2, Papuan4) (11488)
(Papuan3, Papuan) (10729)
(Papuan1, Papuan) (10622)
(Papuan1, Papuan3) (10621)
(Papuan3, Papuan4) (10724)
(Papuan1, Papuan4) (10648)
(Papuan, Papuan4) (10791)
(Papuan1, Papuan2) (10802)
Denhanced(Papuan, Papuan; Denisova, Chimp)

Figure S9. D statistics for characterizing the excess 'unknown' ancestry in USR1. **a.** We compared enhanced D-statistics of the form *D(Native American, H2; Papuan, Outgroup)* (test for Australasian admixture) and *D(Native American, H2; Denisovan, Outgroup)* (test for Denisovan admixture), for the Anzick1, Aymara, Surui and USR1 single genomes. On the assumption that Denisovan admixture is driving the apparent excess Australasian ancestry in USR1, we expect both tests to have the same sign and to be similar in magnitude (SI Section 11.4). For reference, we show dotted lines for *x=y* and *D=0* in both axes. The sets of outgroup genomes for the *(H1, H2)* clade that were used for ascertaining the sites for each test are specified in the lateral panels. **b.** We computed *D(Aymara, USR1; H3, Han)* to determine the best available proxy for the 'unknown' ancestry in USR1. In this case, we expect *D* to increase with the genetic relatedness between *H3* and the true 'introgressing population' (SI Section 11.5). **c.** To investigate whether the observed pattern in USR1 could arise by randomly sampling an individual from a population with already variable Denisovan admixture proportions, we computed the enhanced *D* statistic *D(Papuan, Papuan; Denisova, Chimp)* on sites where the five African genomes and the chimp carry the same allele, for all 10 possible pairs of Papuan genomes (SI Section 11.6). For all tests, one allele from each one of the four genomes was randomly sampled and transition SNPs were excluded. In b and c, the total number of 'ABBA+BABA' sites is shown in brackets. A schematic representation of the null hypothesis for the Z-test is shown in the middle (*D=0*) and the two possible outcomes of the alternative hypothesis are shown to the left (D<0) and the right (D>0). These schematics in c also show the ascertainment step, where sites are chosen such that the green lineages carry an allele different to that in the pink lineage. For all panels, points represent *D* statistics and error bars represent one and *~3.3* standard errors (which corresponds to a *p*-value of ~0.001).

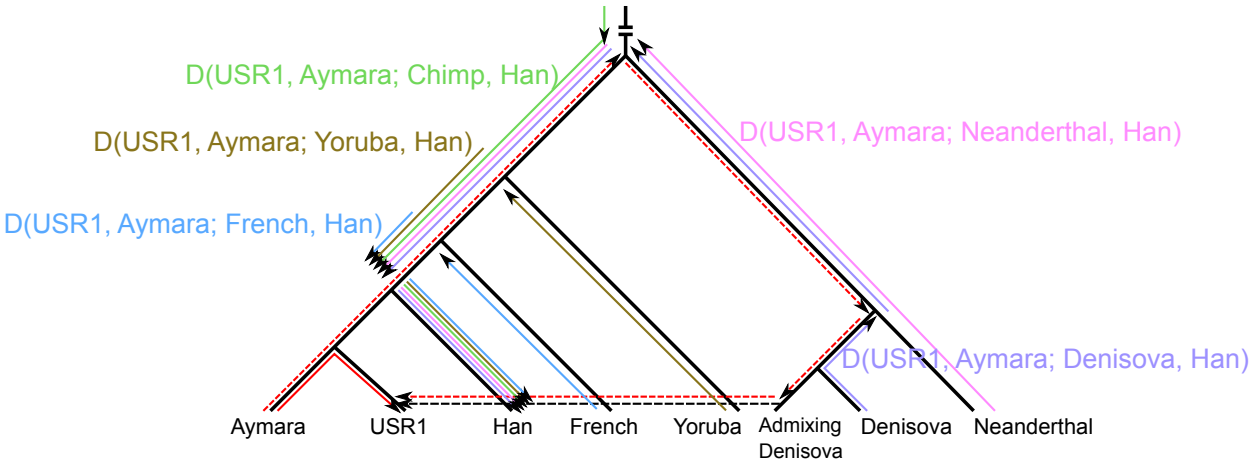

D(USR1, Aymara; Chimp, Han)
D(USR1, Aymara; Yoruba, Han)
D(USR1, Aymara; French, Han)
D(USR1, Aymara; Neanderthal, Han)
D(USR1, Aymara; Denisova, Han)
Aymara
USR1
Han
French
Yoruba
Admixing Denisova
Denisova
Neanderthal

Figure S10. Schematic representation of the expected *D* statistics assuming USR1 carries excess Denisovan admixture compared to Aymara. In SI Section 11.5 we compute D statistics of the form *D(Aymara, USR1; H3, Han)*, where *H3* represents the French, Papuan, Yoruba, Altai Neandertal, Denisovan and chimp genomes. Based on the fact that *H3*=Denisova yields the highest value of *D,* we concluded that the Denisovan genome is the best proxy for the non-Native American ancestry component in USR1. We show a diagram based on [75], where drift paths (colored arrows) are plotted over a tree where USR1 carries excess Denisovan ancestry compared to the Aymara genome. Red arrows represent the Aymara→USR1 drift path, where the solid arrow goes through the common ancestor of Aymara and USR1, and the dotted arrow goes through the Denisovan→USR1 admixture edge (black dotted arrow). The rest of the solid arrows represent *H3*→Han drift paths. Assuming that USR1 carries a Denisovan ancestry component, *D(Aymara, USR1; H3, Han)* is proportional to the overlap between the dotted red arrow and the solid arrows. Note that in the absence of admixture, the colored and red paths would not overlap, thus *D* would be consistent with 0. The relative magnitude of these overlaps (allele frequency difference correlation) is consistent with the *D* statistics shown in Figure S9b.

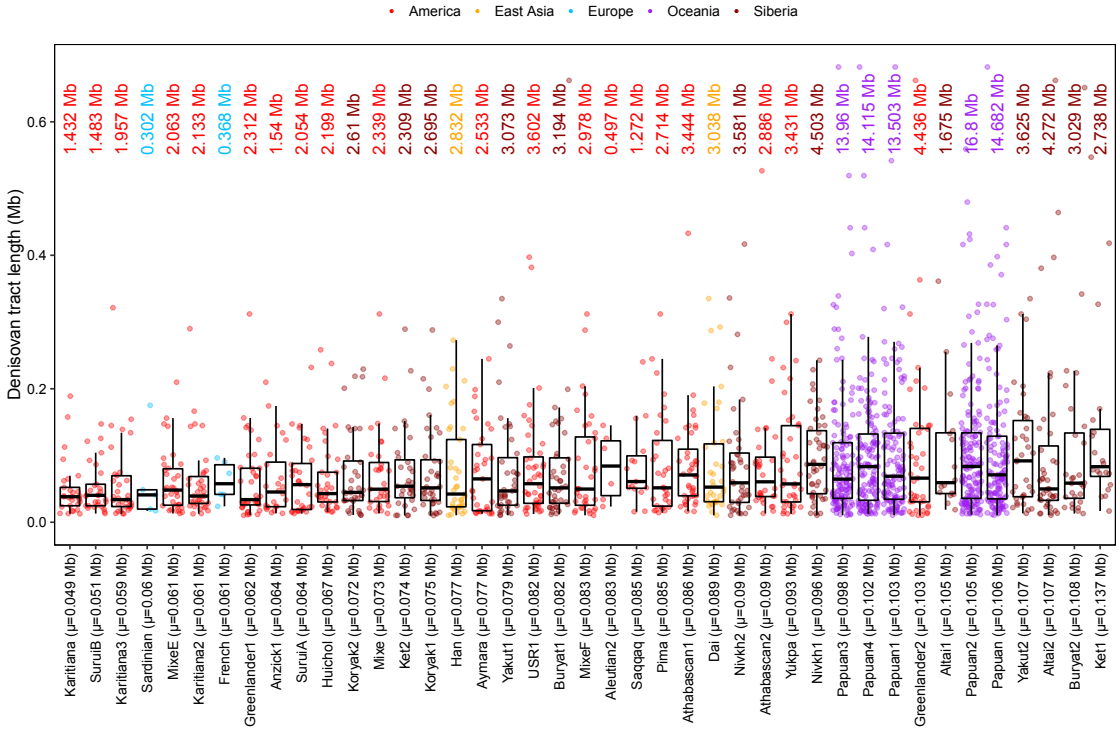

America
East Asia
Europe
Oceania
Siberia
Denisovan tract length (Mb)
0.0
0.2
0.4
0.6
1.432 Mb
1.483 Mb
1.957 Mb
0.302 Mb
2.063 Mb
2.133 Mb
0.368 Mb
2.312 Mb
1.54 Mb
2.054 Mb
2.199 Mb
2.61 Mb
2.339 Mb
2.309 Mb
2.695 Mb
2.832 Mb
2.533 Mb
3.073 Mb
3.602 Mb
3.194 Mb
2.978 Mb
0.497 Mb
1.272 Mb
2.714 Mb
3.444 Mb
3.038 Mb
3.581 Mb
2.886 Mb
3.431 Mb
4.503 Mb
13.96 Mb
14.115 Mb
13.503 Mb
4.436 Mb
1.675 Mb
16.8 Mb
14.682 Mb
3.625 Mb
4.272 Mb
3.029 Mb
2.738 Mb
Karitiana (μ=0.049 Mb)
SuruiB (μ=0.051 Mb)
Karitiana3 (μ=0.059 Mb)
Sardinian (μ=0.06 Mb)
MixeE (μ=0.061 Mb)
Karitiana2 (μ=0.061 Mb)
French (μ=0.061 Mb)
Greenlander1 (μ=0.062 Mb)
Anzick1 (μ=0.064 Mb)
SuruiA (μ=0.064 Mb)
Huichol (μ=0.067 Mb)
Koryak2 (μ=0.072 Mb)
Mixe (μ=0.073 Mb)
Ket2 (μ=0.074 Mb)
Koryak1 (μ=0.075 Mb)
Han (μ=0.077 Mb)
Aymara (μ=0.077 Mb)
Yakut1 (μ=0.079 Mb)
USR1 (μ=0.082 Mb)
Buryat1 (μ=0.082 Mb)
MixeF (μ=0.083 Mb)
Aleutian2 (μ=0.083 Mb)
Saqqaq (μ=0.085 Mb)
Pima (μ=0.085 Mb)
Athabascan1 (μ=0.086 Mb)
Dai (μ=0.089 Mb)
Nivkh2 (μ=0.09 Mb)
Athabascan2 (μ=0.09 Mb)
Yukpa (μ=0.093 Mb)
Nivkh1 (μ=0.096 Mb)
Papuan3 (μ=0.098 Mb)
Papuan4 (μ=0.102 Mb)
Papuan1 (μ=0.103 Mb)
Greenlander2 (μ=0.103 Mb)
Altai1 (μ=0.105 Mb)
Papuan2 (μ=0.105 Mb)
Papuan (μ=0.106 Mb)
Yakut2 (μ=0.107 Mb)
Altai2 (μ=0.107 Mb)
Buryat2 (μ=0.108 Mb)
Ket1 (μ=0.137 Mb)

Figure S11. Denisovan ancestry tract length distributions in high coverage genomes. We inferred Denisovan ancestry tracts in USR1 and the high-coverage reference genomes. For each genome, we show the lengths (in Mb) of individual tracts and a boxplot of their distribution. Boxes show the median and are delimited by the first and third quartiles. Whiskers represent the largest value within 1.5 times the inter-quartile range, starting from their respective end of the box. Mean tract lengths are indicated in brackets, and the sum of the lengths of all tracts is displayed in color for each genome. For each single genome, we obtained the following counts of Denisovan ancestry tracts: Aleutian2: 6, Altai1: 16, Altai2: 40, Anzick1: 24, Athabascan1: 40, Athabascan2: 32, Aymara: 33, Buryat1: 39, Buryat2: 28, Dai: 34, French: 6, Greenlander1: 37, Greenlander2: 43, Han: 37, Huichol: 33, Karitiana: 29, Karitiana2: 35, Karitiana3: 33, Ket1: 20, Ket2: 31, Koryak1: 36, Koryak2: 36, Mixe: 32, MixeE: 34, MixeF: 36, Nivkh1: 47, Nivkh2: 40, Papuan: 138, Papuan1: 131, Papuan2: 160, Papuan3: 142, Papuan4: 138, Pima: 32, Saqqaq: 15, Sardinian: 5, SuruiA: 32, SuruiB: 29, USR1: 44, Yakut1: 39, Yakut2: 34, Yukpa: 37.

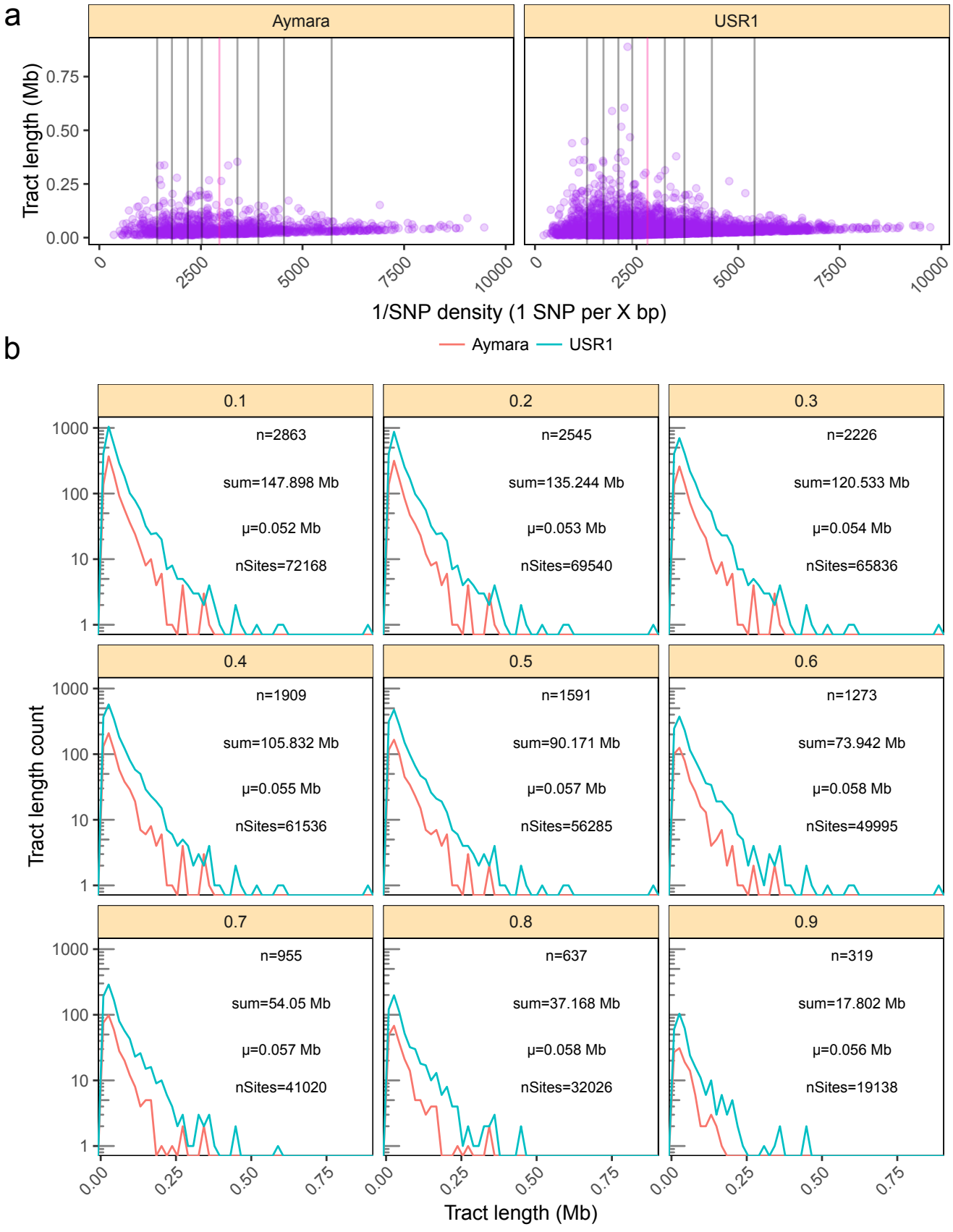

a
Aymara
USR1
Tract length (Mb)
0.00
0.25
0.50
0.75
0
2500
5000
7500
10000
1/SNP density (1 SNP per X bp)
b
Aymara
USR1
0.1
n=2863
sum=147.898 Mb
μ=0.052 Mb
nSites=72168
0.2
n=2545
sum=135.244 Mb
μ=0.053 Mb
nSites=69540
0.3
n=2226
sum=120.533 Mb
μ=0.054 Mb
nSites=65836
0.4
n=1909
sum=105.832 Mb
μ=0.055 Mb
nSites=61536
0.5
n=1591
sum=90.171 Mb
μ=0.057 Mb
nSites=56285
0.6
n=1273
sum=73.942 Mb
μ=0.058 Mb
nSites=49995
0.7
n=955
sum=54.05 Mb
μ=0.057 Mb
nSites=41020
0.8
n=637
sum=37.168 Mb
μ=0.058 Mb
nSites=32026
0.9
n=319
sum=17.802 Mb
μ=0.056 Mb
nSites=19138
Tract length count
1
10
100
1000
0.00
0.25
0.50
0.75
Tract length (Mb)

Figure S12. Characterization of USR1- and Aymara-specific tracts. We defined USR1- and Aymara-specific tracts, based on sites for which all Native American genomes (SI Section 6) are fixed and USR1 or Aymara carry a different allele, respectively. **a.** For each tract, we computed the '*X*-specific' site density, defined as the number of '*X*-specific' sites in the tract divided by the tract length. We found 1069 Aymara-specific tracts and 3181 USR1-specific tracts. We show the tract lengths (in Mb) as a function of the inverse of the '*X*-specific' site density. Tract sets were defined according to the deciles (10-90%) of the density distribution (SI Section 13), depicted by vertical black lines. The red line corresponds to the 50% decile; 'USR1-specific' tracts with a density above this decile were masked for subsequent tests (SI Section 13.3). **b.** Tract length distribution for 'dense' tract sets. For each decile (header panels) of the density distribution, only the tracts with a density above the decile were considered. In each panel, we show the number of tracts with a density above the decile, the sum of their lengths, their mean length, and the number of 'X-specific' sites encompassed by these tracts.

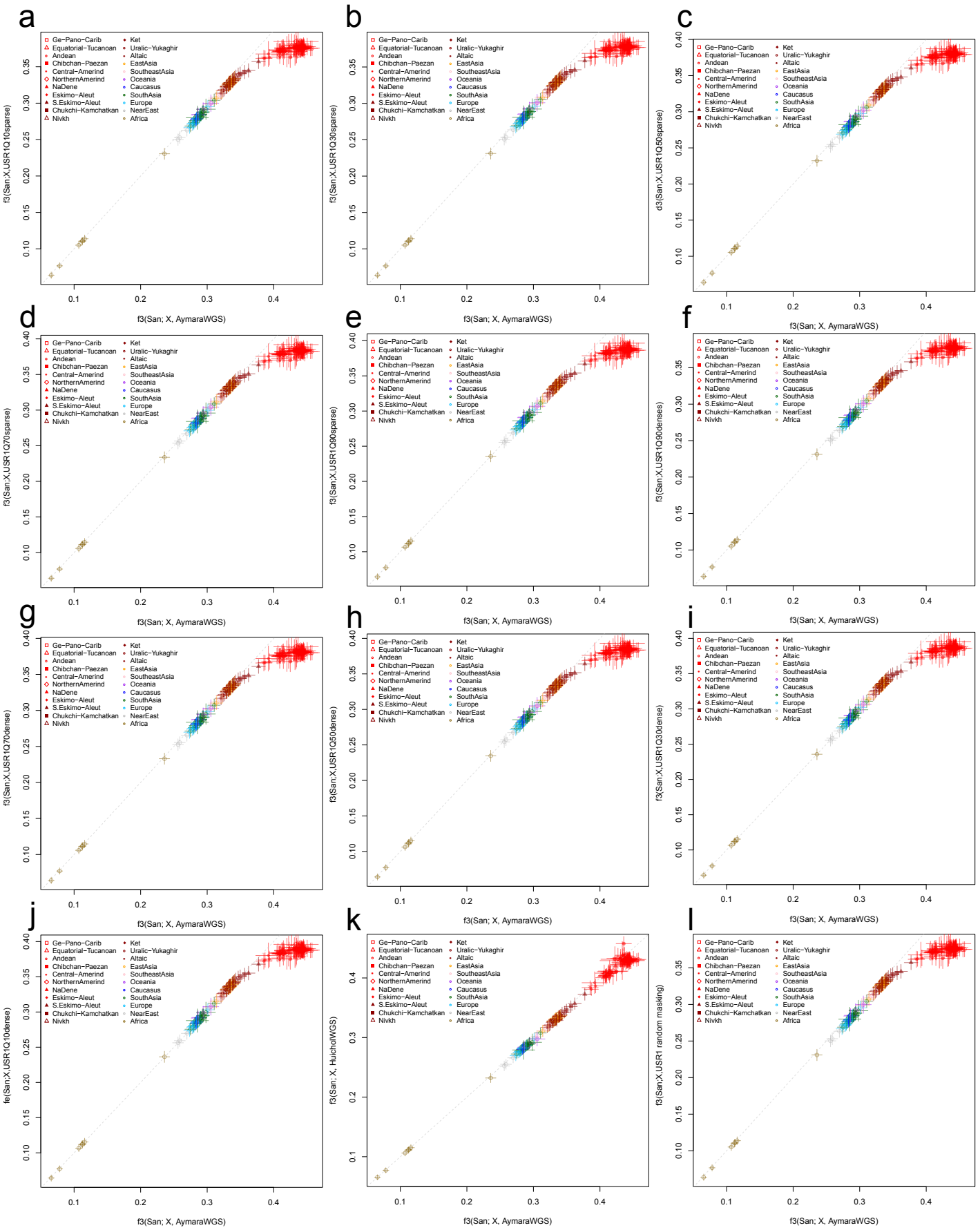

a
b
c
d
e
f
g
h
i
j
k
l
Ge−Pano−Carib
Equatorial−Tucanoan
Andean
Chibchan−Paezan
Central−Amerind
NorthernAmerind
NaDene
Eskimo−Aleut
S.Eskimo−Aleut
Chukchi−Kamchatkan
Nivkh
Ket
Uralic−Yukaghir
Altaic
EastAsia
SoutheastAsia
Oceania
Caucasus
SouthAsia
Europe
NearEast
Africa
f3(San;X,USR1Q10sparse)
f3(San;X,USR1Q30sparse)
d3(San;X,USR1Q50sparse)
f3(San;X,USR1Q70sparse)
f3(San;X,USR1Q90sparse)
f3(San;X,USR1Q90dense)
f3(San;X,USR1Q70dense)
f3(San;X,USR1Q50dense)
f3(San;X,USR1Q30dense)
fe(San;X,USR1Q10dense)
f3(San; X, HulchoWGS)
f3(San;X,USR1 random masking)
f3(San; X, AymaraWGS)

Figure S13. Pairwise outgroup f3 statistics after masking USR1-specific tracts. We masked the USR1 genome according to the different density-based tract sets (SI Section 13), computed *f3(San; USR1, X)* and compared these results to those obtained for *f3(San; Aymara, X)* (SI Section 9*).* **a-e**. Pairwise *f3* statistics after masking 'sparse' tracts with density BELOW the 10%, 30%, 50%, 70% and 90% deciles of the 'X-specific' site density distribution, respectively. **f-j**. Pairwise *f3* statistics after masking 'dense' tracts with density ABOVE the 90%, 70%, 50%, 30% and 10% deciles of the 'X-specific' site density distribution. **k.** Pairwise comparison of *f3(San; Huichol, X) and f3(San; Aymara, X)*. In the Mexican Huichol genome, we observe a pattern similar to that observed in USR1, in which the overall statistics are slightly higher for the Aymara genome. **l.** Pairwise *f3* statistics comparison after masking random tracts in the USR1 genome, with a tract length distribution similar to that of the 'dense' tracts with density greater than the 50% decile of the density distribution. For all tests, we used the genotype panel described in Section 6.1. Points show pairwise *f3* statistics, and error bars show ~2.57 standard errors, which corresponds to a p-value of 0.01 in a Z-test.

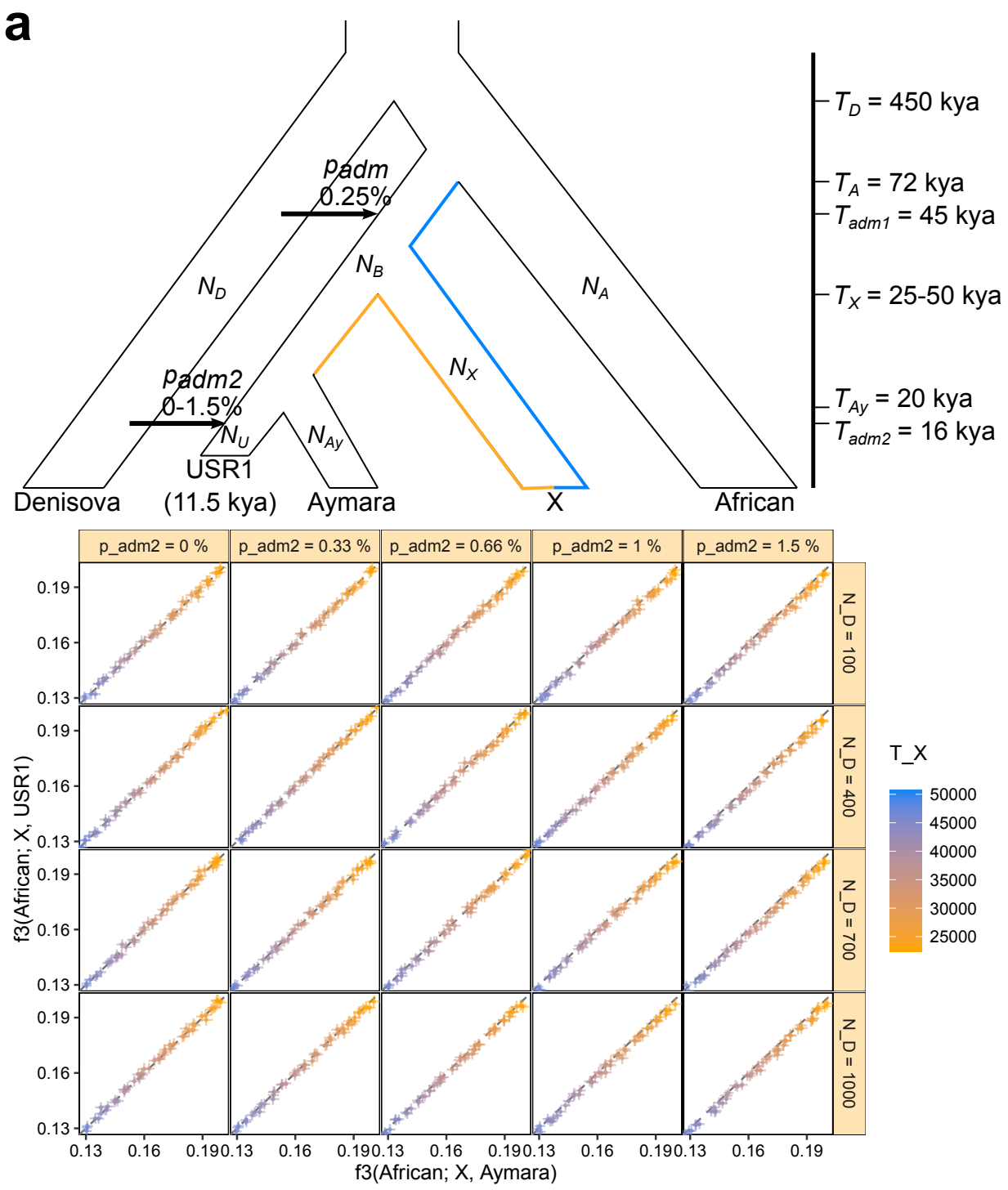

a
T_D = 450 kya
T_A = 72 kya
T_{adm1} = 45 kya
T_X = 25-50 kya
T_{Ay} = 20 kya
T_{adm2} = 16 kya
padm
0.25%
padm2
0-1.5%
N_D
N_B
N_A
N_X
N_U
N_{Ay}
Denisova
USR1
(11.5 kya)
Aymara
X
African
p_adm2 = 0 %
p_adm2 = 0.33 %
p_adm2 = 0.66 %
p_adm2 = 1 %
p_adm2 = 1.5 %
N_D = 100
N_D = 400
N_D = 700
N_D = 1000
f3(African; X, USR1)
f3(African; X, Aymara)
0.13
0.16
0.19
T_X
50000
45000
40000
35000
30000
25000

Figure S14. Simulation study of Denisovan ancestry in USR1. a. Schematic representation of the demographic model that was used for the coalescent simulations detailed in SI Section 13.4. We assumed the population split and admixture times shown in the right and the following population sizes: N_D={100, 400, 700, 1000}, N_A=20,000, N_B=1,600, N_{Ay}=1,500, N_U=2,000 and N_X={3000, 8000, 13000}. Note that different values of T_X correspond to different splitting times in the branch shown in color, and that T_X can take values that pre- and post-date T_{adm2}. For each simulation, we sampled ten diploid individuals from the African and X populations, one diploid individual from the Aymara population and one diploid individual from the USR1 population, which was sampled 11.5 kya. We considered two admixture events from Denisovans: one into the out of Africa population (p_{adm}=0.0025) and one into USR1 (p_{adm2}={0, 0.0033, 0.0066, 0.01, 0.015}). **b.** We computed *f3(African; X, USR1)* and *f3(African; X, Aymara)* for each simulated dataset, for different instances of population X (Eurasians). Different panels correspond to different combinations of the simulated population size for Denisovans (N_D), and the admixture proportion into USR1 (p_{adm2}). The color scheme corresponds to the T_X parameter. Points show pairwise *f3* statistics and error bars show ~2.57 standard errors, which corresponds to a *p*-value of 0.01 in a *Z*-test.

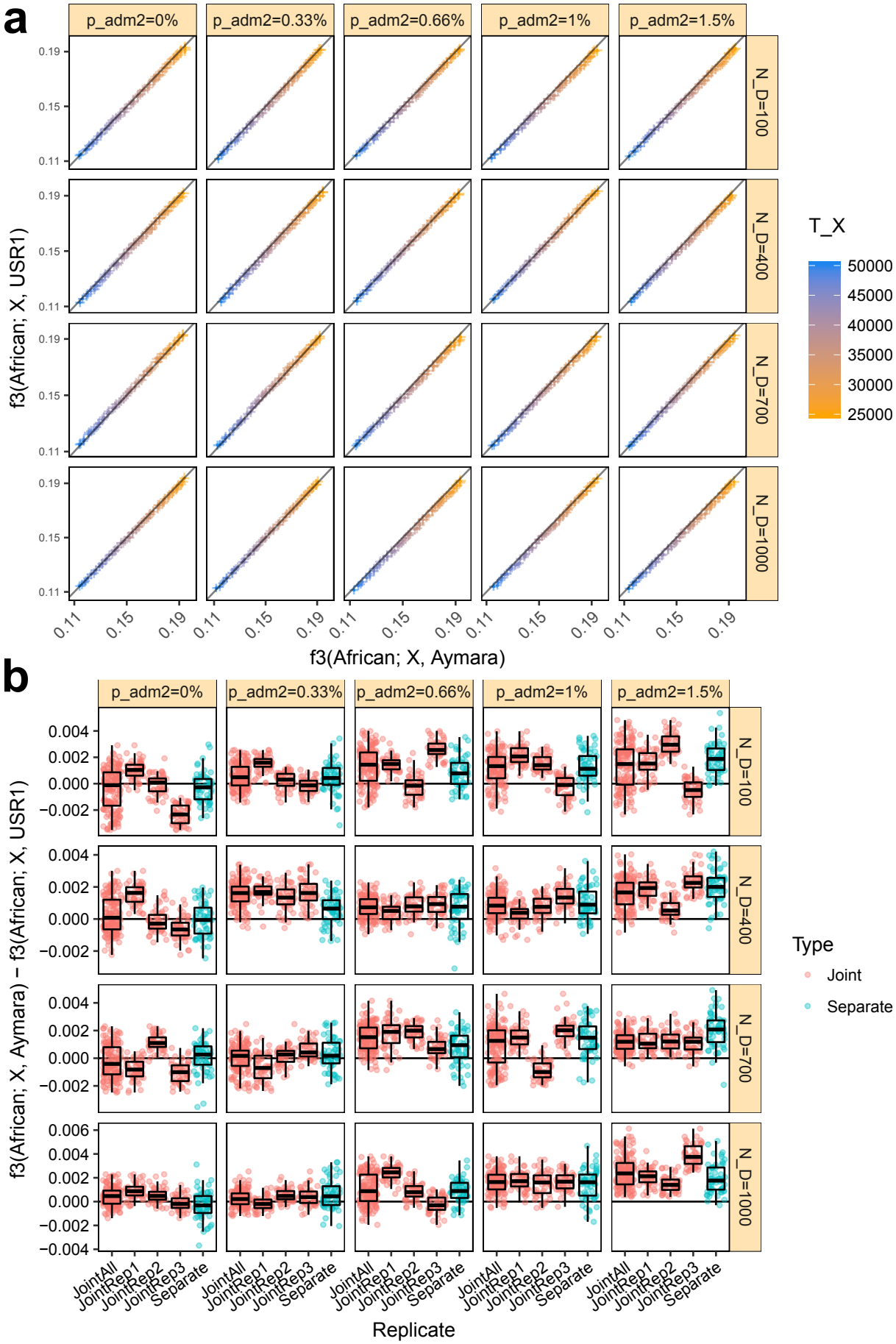
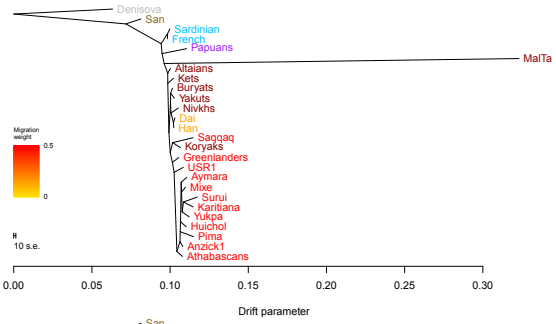
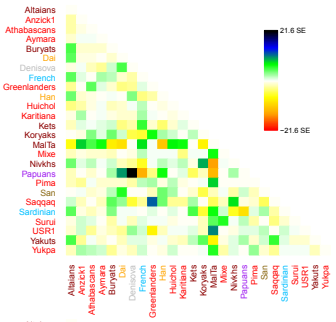
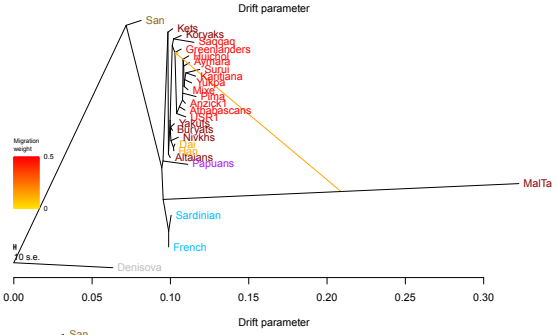
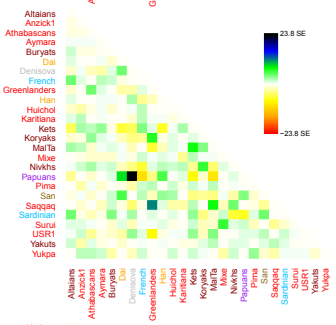

a
p_adm2=0%
p_adm2=0.33%
p_adm2=0.66%
p_adm2=1%
p_adm2=1.5%
N_D=100
N_D=400
N_D=700
N_D=1000
f3(African; X, USR1)
f3(African; X, Aymara)
T_X
50000
45000
40000
35000
30000
25000
b
f3(African; X, Aymara) − f3(African; X, USR1)
Replicate
JointAll
JointRep1
JointRep2
JointRep3
Separate
Type
Joint
Separate

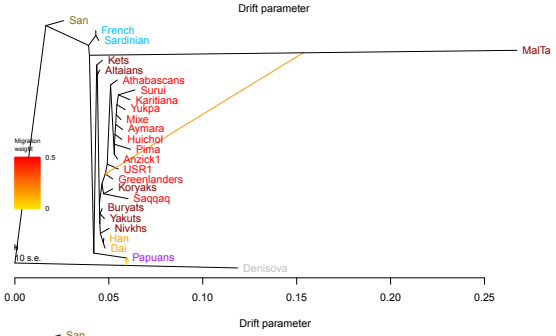
Figure S15. Alternative setup for the simulation study of Denisovan ancestry in USR1. a. We carried out a simulation study similar to that detailed in Section 13.4, using the same parameters. The difference is that in this case, we jointly simulated the 'African' population, the single 'USR1' and 'Aymara' individuals, and 51 instances of *X*. The latter were assigned random effective population sizes (N_{Xi}) between 3,000 and 13,000 and split times (in years) T_{Xi}={25,000, 25,500, ..., 50,000}. Thus, for each panel only one 'USR1' and one 'Aymara' genome were simulated for all instances of *X,* in contrast to the model shown in Figure S14b, where each point represents an independently simulated USR1 and Aymara genomes. This figure shows the results for one replicate of the alternative simulation setup. The color scheme corresponds to the T_{Xi} parameter, points show pairwise *f3* statistics and error bars correspond to ~2.57 standard errors (*p*-value ~ 0.01). **b.** Distribution of the size of the pairwise *f3* offset for different simulation setups. In addition to the results shown in Figure S14b (here termed 'Separate'), we simulated three replicates of a demographic scenario where all instances of population *X* are simulated jointly (here termed 'Joint'). Additionally, we combined the three replicates of the 'Joint' setup (here termed 'JointAll'). For each instance of *X*, for each of the simulation setups, we computed the size of the *f3*-statistic offset, by computing *f3(African; X, Aymara)-f3(African; X, USR1)*. We display boxplots that represent the distribution of these differences, for each combination of p_{adm2} and N_D. Boxes show the median and are delimited by the first and third quartiles. Whiskers represent the largest value within 1.5 times the inter-quartile range, starting from their respective end of the box.

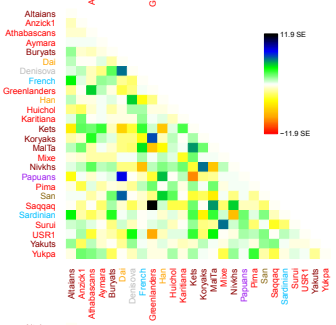

Denisova
San
Sardinian
French
Papuans
MalTa
Altaians
Kets
Buryats
Yakuts
Nivkhs
Dai
Han
Saqqaq
Koryaks
Greenlanders
USR1
Aymara
Mixe
Surui
Karitiana
Yukpa
Huichol
Pima
Anzick1
Athabascans
Migration weight
0.5
0
10 s.e.
0.00
0.05
0.10
0.15
0.20
0.25
0.30
Drift parameter

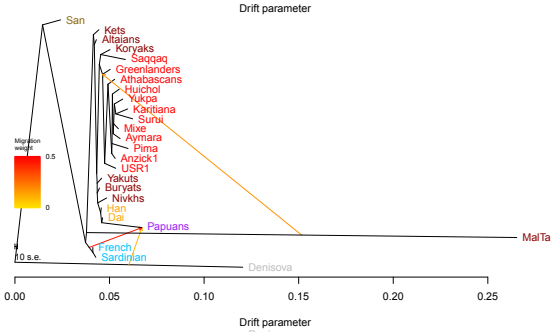

21.6 SE
-21.6 SE

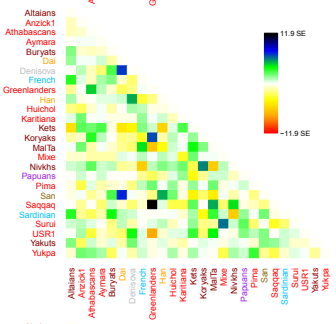

San
Kets
Koryaks
Saqqaq
Greenlanders
Huichol
Aymara
Surui
Karitiana
Yukpa
Mixe
Pima
Anzick1
Athabascans
USR1
Buryats
Nivkhs
Dai
Han
Altaians
Papuans
MalTa
Sardinian
French
Denisova
Migration weight
0.5
0
10 s.e.
0.00
0.05
0.10
0.15
0.20
0.25
0.30
Drift parameter

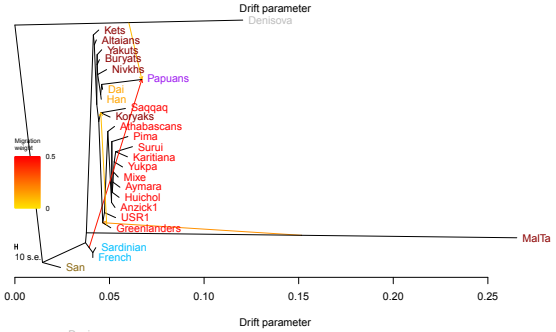

23.8 SE
-23.8 SE

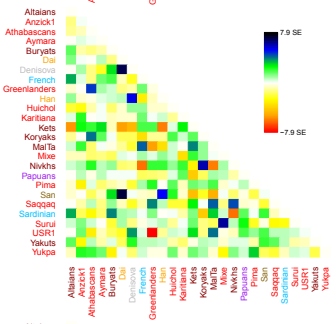

San
French
Sardinian
MalTa
Kets
Altaians
Athabascans
Surui
Karitiana
Yukpa
Mixe
Aymara
Huichol
Pima
Anzick1
USR1
Greenlanders
Koryaks
Saqqaq
Buryats
Yakuts
Nivkhs
Han
Dai
Papuans
Denisova
Migration weight
0.5
0
10 s.e.
0.00
0.05
0.10
0.15
0.20
0.25
Drift parameter

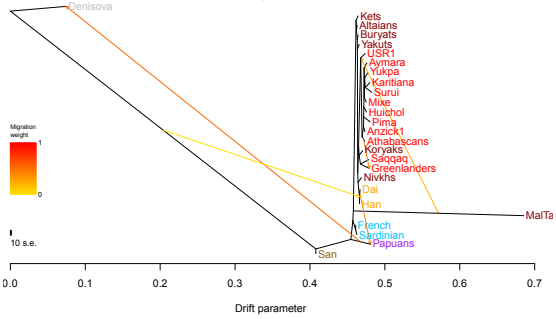

11.9 SE
-11.9 SE


San
Kets
Altaians
Koryaks
Saqqaq
Greenlanders
Athabascans
Huichol
Yukpa
Karitiana
Surui
Mixe
Aymara
Pima
Anzick1
USR1
Yakuts
Buryats
Nivkhs
Han
Dai
Papuans
MalTa
French
Sardinian
Denisova
Migration weight
0.5
0
10 s.e.
0.00
0.05
0.10
0.15
0.20
0.25
Drift parameter


11.9 SE
-11.9 SE


Denisova
Kets
Altaians
Yakuts
Buryats
Nivkhs
Papuans
Dai
Han
Saqqaq
Koryaks
Athabascans
Pima
Surui
Karitiana
Yukpa
Mixe
Aymara
Huichol
Anzick1
USR1
Greenlanders
MalTa
Sardinian
French
San
Migration weight
0.5
0
10 s.e.
0.00
0.05
0.10
0.15
0.20
0.25
Drift parameter


7.9 SE
-7.9 SE


Denisova
Kets
Altaians
Buryats
Yakuts
USR1
Aymara
Yukpa
Karitiana
Surui
Mixe
Huichol
Pima
Anzick1
Athabascans
Koryaks
Saqqaq
Greenlanders
Nivkhs
Dai
Han
MalTa
French
Sardinian
Papuans
San
Migration weight
1
0
10 s.e.
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
Drift parameter

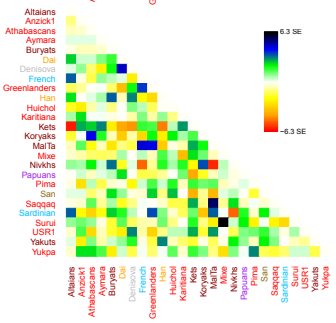

6.3 SE
-6.3 SE

Figure S16. Admixture graphs based on modern and ancient whole genomes. We fitted admixture graphs with zero to five admixture edges to the whole-genome dataset. After excluding transition SNPs, we considered 631,192 variable sites (SI Section 6). On the left, we show the admixture graphs produced by *TreeMix* for increasing number of admixture edges. On the right, we show the pairwise residuals after fitting the maximum likelihood graph.

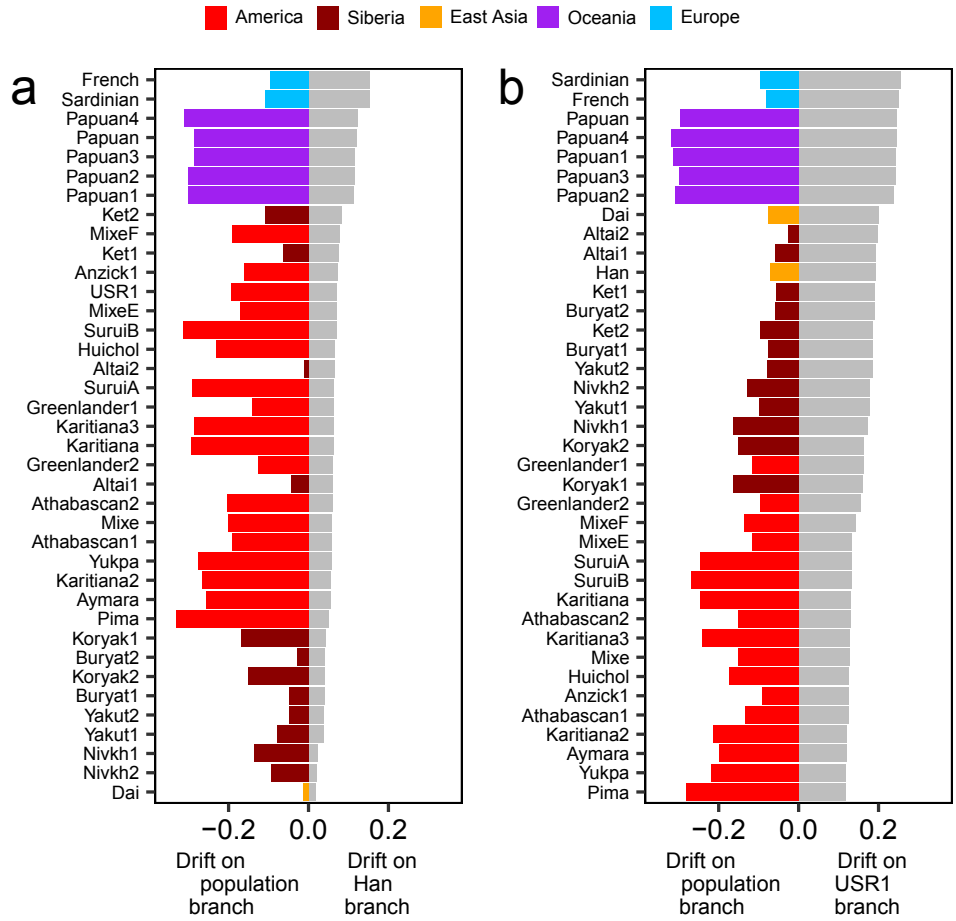

America
Siberia
East Asia
Oceania
Europe
a
French
Sardinian
Papuan4
Papuan
Papuan3
Papuan2
Papuan1
Ket2
MixeF
Ket1
Anzick1
USR1
MixeE
SuruiB
Huichol
Altai2
SuruiA
Greenlander1
Karitiana3
Karitiana
Greenlander2
Altai1
Athabascan2
Mixe
Athabascan1
Yukpa
Karitiana2
Aymara
Pima
Koryak1
Buryat2
Koryak2
Buryat1
Yakut2
Yakut1
Nivkh1
Nivkh2
Dai
−0.2
0.0
0.2
Drift on population branch
Drift on Han branch
b
Sardinian
French
Papuan
Papuan4
Papuan1
Papuan3
Papuan2
Dai
Altai2
Altai1
Han
Ket1
Buryat2
Ket2
Buryat1
Yakut2
Nivkh2
Yakut1
Nivkh1
Koryak2
Greenlander1
Koryak1
Greenlander2
MixeF
MixeE
SuruiA
SuruiB
Karitiana
Athabascan2
Karitiana3
Mixe
Huichol
Anzick1
Athabascan1
Karitiana2
Aymara
Yukpa
Pima
−0.2
0.0
0.2
Drift on population branch
Drift on USR1 branch

Figure S17. Drift after divergence for a set of high coverage genomes. We used the method originally described in [44], to obtain maximum likelihood estimates (MLEs) for the amount of drift after the split for pairs of whole genomes. We obtained MLEs for all pairs involving a high-coverage genome and a fixed a 'test' genome. For each comparison, drift in the lineage leading to the fixed 'test' genome is displayed in gray. **a.** MLEs for all pairs including the Han Chinese genome. After splitting from Native Americans, the drift in the branch that leads to the Han genome is similar for all tests. **b.** MLEs for all comparisons including the USR1 genome. The USR1 branch has similar drift estimates for all comparisons involving a Native American genome.

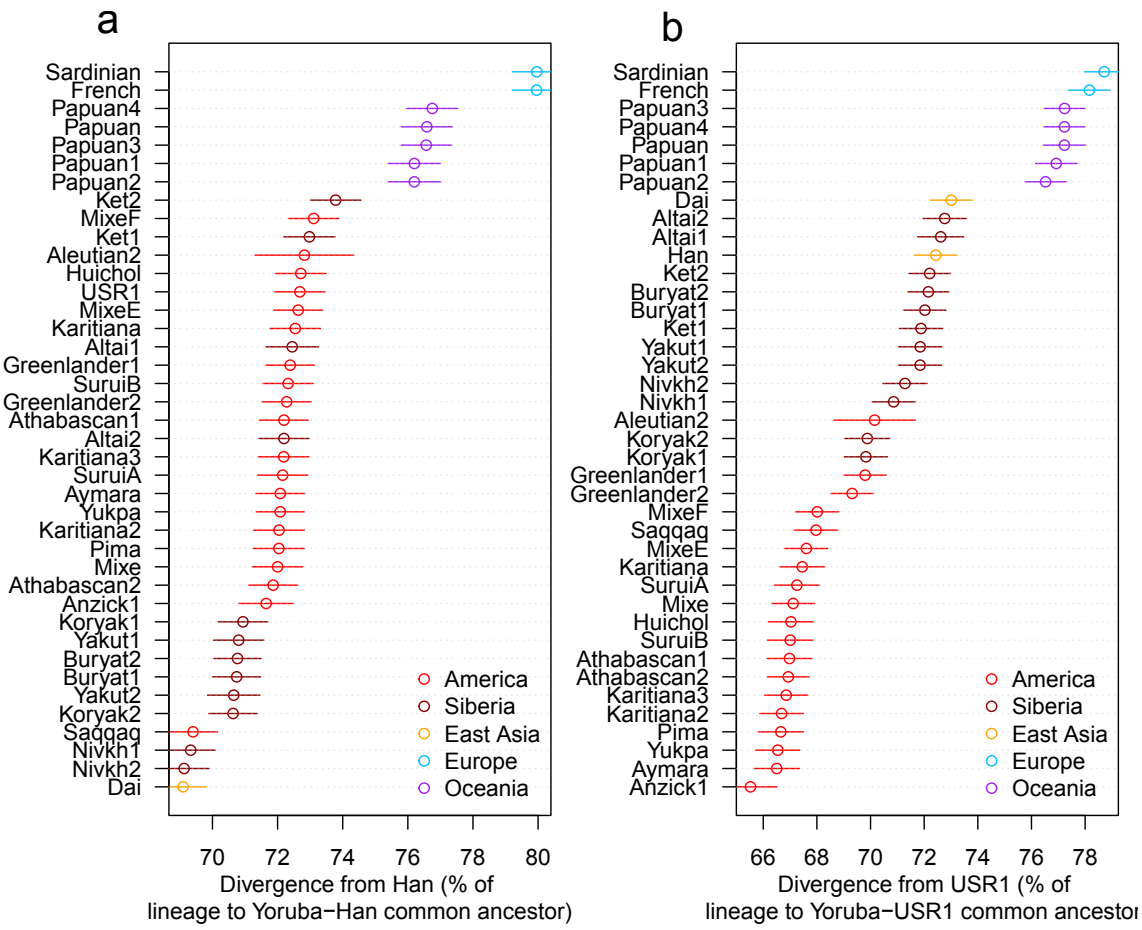

a
Sardinian
French
Papuan4
Papuan
Papuan3
Papuan1
Papuan2
Ket2
MixeF
Ket1
Aleutian2
Huichol
USR1
MixeE
Karitiana
Altai1
Greenlander1
SuruiB
Greenlander2
Athabascan1
Altai2
Karitiana3
SuruiA
Aymara
Yukpa
Karitiana2
Pima
Mixe
Athabascan2
Anzick1
Koryak1
Yakut1
Buryat2
Buryat1
Yakut2
Koryak2
Saqqaq
Nivkh1
Nivkh2
Dai
America
Siberia
East Asia
Europe
Oceania
70
72
74
76
78
80
Divergence from Han (% of
lineage to Yoruba−Han common ancestor)
b
Sardinian
French
Papuan3
Papuan4
Papuan
Papuan1
Papuan2
Dai
Altai2
Altai1
Han
Ket2
Buryat2
Buryat1
Ket1
Yakut1
Yakut2
Nivkh2
Nivkh1
Aleutian2
Koryak2
Koryak1
Greenlander1
Greenlander2
MixeF
Saqqaq
MixeE
Karitiana
SuruiA
Mixe
Huichol
SuruiB
Athabascan1
Athabascan2
Karitiana3
Karitiana2
Pima
Yukpa
Aymara
Anzick1
America
Siberia
East Asia
Europe
Oceania
66
68
70
72
74
76
78
Divergence from USR1 (% of
lineage to Yoruba−USR1 common ancestor

Figure S18. Pairwise average genomic divergence for a set of high coverage genomes. We show the average divergence between a fixed 'test' genome and a set of high-coverage genomes. **a.** Average divergence between the Han Chinese genome and a worldwide sample of high-coverage genomes. The Han genome is similarly diverged from all Native American genomes. **b.** Average divergence between the USR1 genome and the high-coverage genomes. All estimates are relative to the branch that leads from the common ancestor of the genome pair and an African outgroup (Yoruba) to the test genome. For each test, points represent pairwise average genomic divergence and we show a 95% confidence interval, based on the standard errors from a block jackknife procedure.

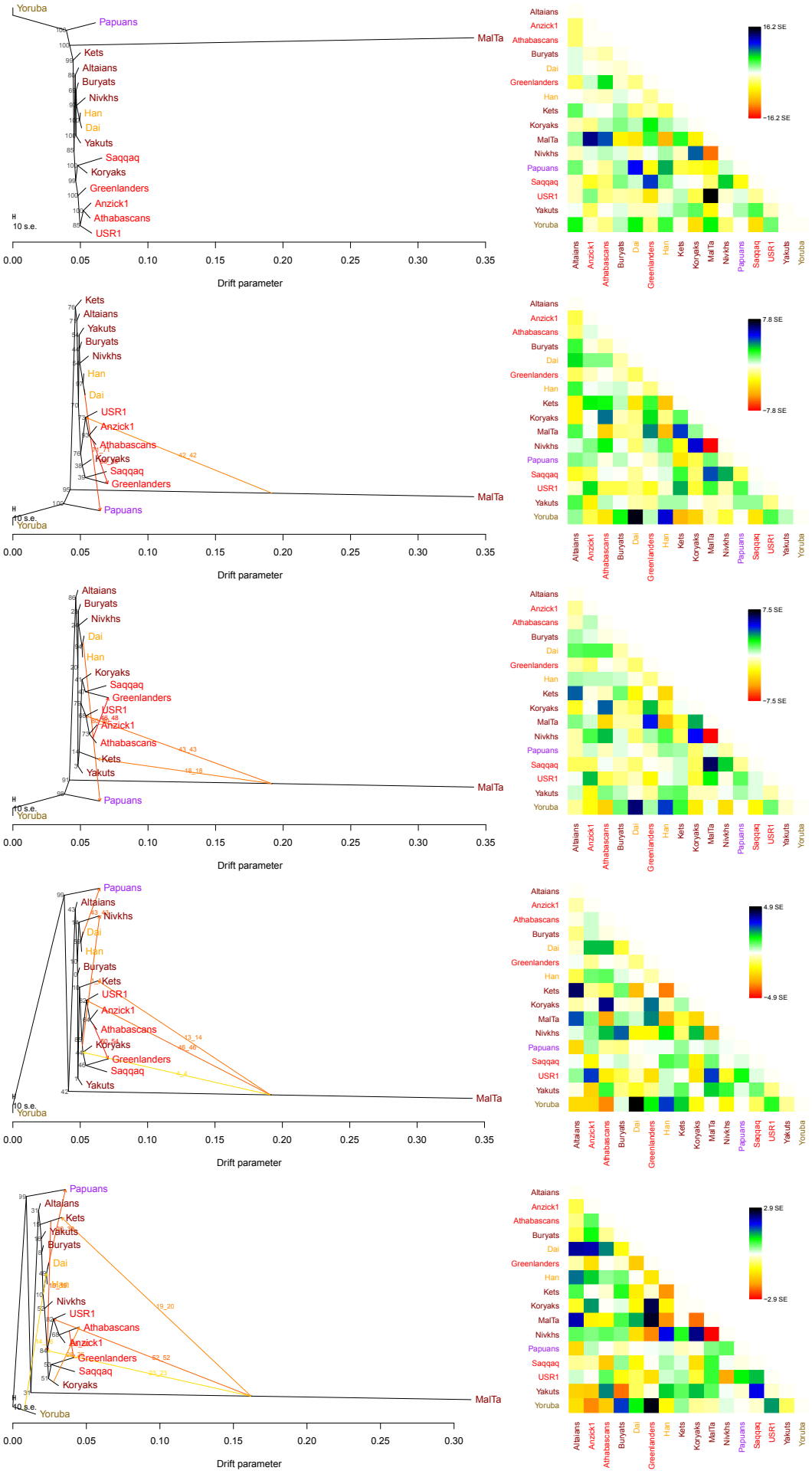

Yoruba
Papuans
MalTa
Kets
Altaians
Buryats
Nivkhs
Han
Dai
Yakuts
Saqqaq
Koryaks
Greenlanders
Anzick1
Athabascans
USR1
H
10 s.e.
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
Drift parameter
16.2 SE
−16.2 SE
7.8 SE
−7.8 SE
7.5 SE
−7.5 SE
4.9 SE
−4.9 SE
2.9 SE
−2.9 SE

Figure S19. Admixture graphs exploring the origin of Inuit and Athabascans. We fitted admixture graphs with incremental numbers of admixture edges until we observed gene flow into both admixed Native American groups (SI Section 17). After excluding transition SNPs, we considered 575,141 variable sites (SI Section 6). On the left, we show the admixture graphs produced by *TreeMix* for 0,3,4,6 and 8 admixture edges. On the right, we show the pairwise residuals after fitting the maximum likelihood graph. Bootstrap support for each node is shown with black numbers on left of each node (in percentage). Support for each migration edge is shown with color numbers on the right. For the latter, the first number is the support for a given edge in the displayed direction and the second number (after the underscore) is the support for the edge in any direction.

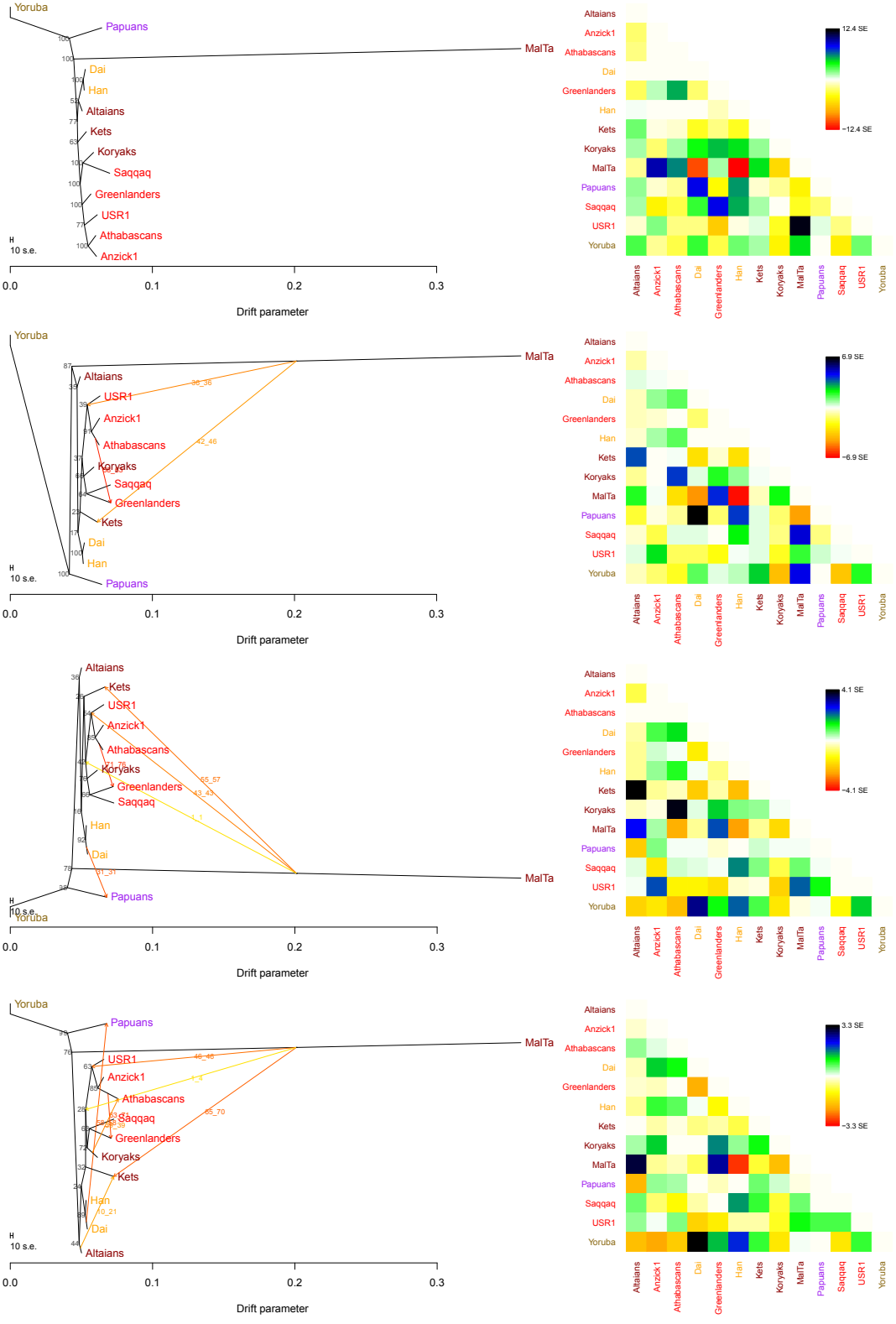

Yoruba
Papuans
MalTa
Dai
Han
Altaians
Kets
Koryaks
Saqqaq
Greenlanders
USR1
Athabascans
Anzick1
H
10 s.e.
0.0
0.1
0.2
0.3
Drift parameter
12.4 SE
−12.4 SE
6.9 SE
−6.9 SE
4.1 SE
−4.1 SE
3.3 SE
−3.3 SE

Figure S20. Admixture graphs exploring the origin of Inuit and Athabascans, using a reduced dataset with less Siberian ingroups. These results are similar to those depicted in Figure S19, but depend on a reduced dataset from which Nivkhs, Yakuts and Buryats were excluded (SI Section 17). After excluding transition SNPs, we considered 556,229 variable sites and show results for 0,3,5 and 7 admixture edges.

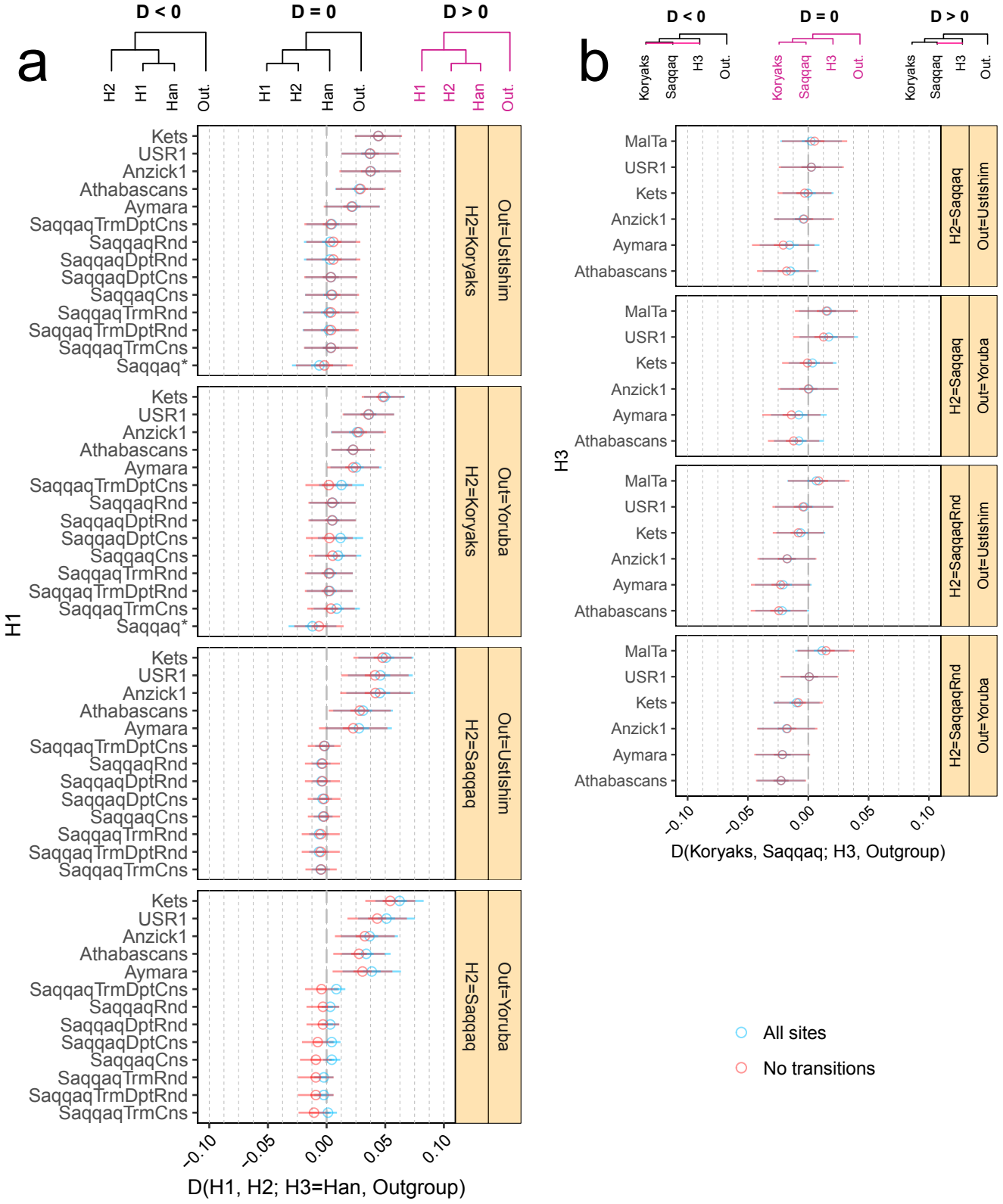

a
D < 0
D = 0
D > 0
H2
H1
Han
Out.
H1
H2
Han
Out.
H1
H2
Han
Out.
Kets
USR1
Anzick1
Athabascans
Aymara
SaqqaqTrmDptCns
SaqqaqRnd
SaqqaqDptRnd
SaqqaqDptCns
SaqqaqCns
SaqqaqTrmRnd
SaqqaqTrmDptRnd
SaqqaqTrmCns
Saqqaq*
H2=Koryaks
Out=Ustlshim
Out=Yoruba
H2=Saqqaq
H1
−0.10
−0.05
0.00
0.05
0.10
D(H1, H2; H3=Han, Outgroup)
b
D < 0
D = 0
D > 0
Koryaks
Saqqaq
H3
Out.
MalTa
USR1
Kets
Anzick1
Aymara
Athabascans
H2=Saqqaq
Out=Ustlshim
Out=Yoruba
H2=SaqqaqRnd
H3
D(Koryaks, Saqqaq; H3, Outgroup)
All sites
No transitions

Figure S21. Pooled whole genome-based *D*-statistics testing the placement of Koryaks and Saqqaq in the admixture graph model. a. We computed D(H1, Koryaks; Han, Outgroup) and (H1, Saqqaq; Han, Outgroup) in order to confirm that Koryaks are the modeled population that bears the largest Asian fraction. Besides the called Saqqaq genotypes, we considered eight alternative Saqqaq versions (SI Section 17.3). Trm: 5 bases were trimmed from the ends of the reads. Dpt: Sites with depth of coverage lower than 5 were excluded. Rnd: A random allele was sampled at each site. Cns: A majority rule consensus was built. **b.** We computed D(Saqqaq, Koryaks; H3, Outgroup) in order to test for bi-directional gene flow between Koryaks and Native Americans. Dendrograms above the plots represent the favored hypothesis depending on the value of *D*, and the expected under the admixture graph model is highlighted in pink. All tests were computed by including all sites (blue) and transversion polymorphisms only (red). Points represent *D* statistics and error bars represent one and ~3.3 standard errors (which corresponds to a *p*-value of *~0.001*). Each Saqqaq alternative version consists of one individual and each 'pooled population' contains the following number of individuals: Yoruba (n=1), Ust'ishim (n=1), Mal'ta (n=1), Han (n=1), Ket (n=2), Koryak (n=2), USR1 (n=1), Anzick1 (n=1), Aymara (n=1) and Athabascan (n=2).

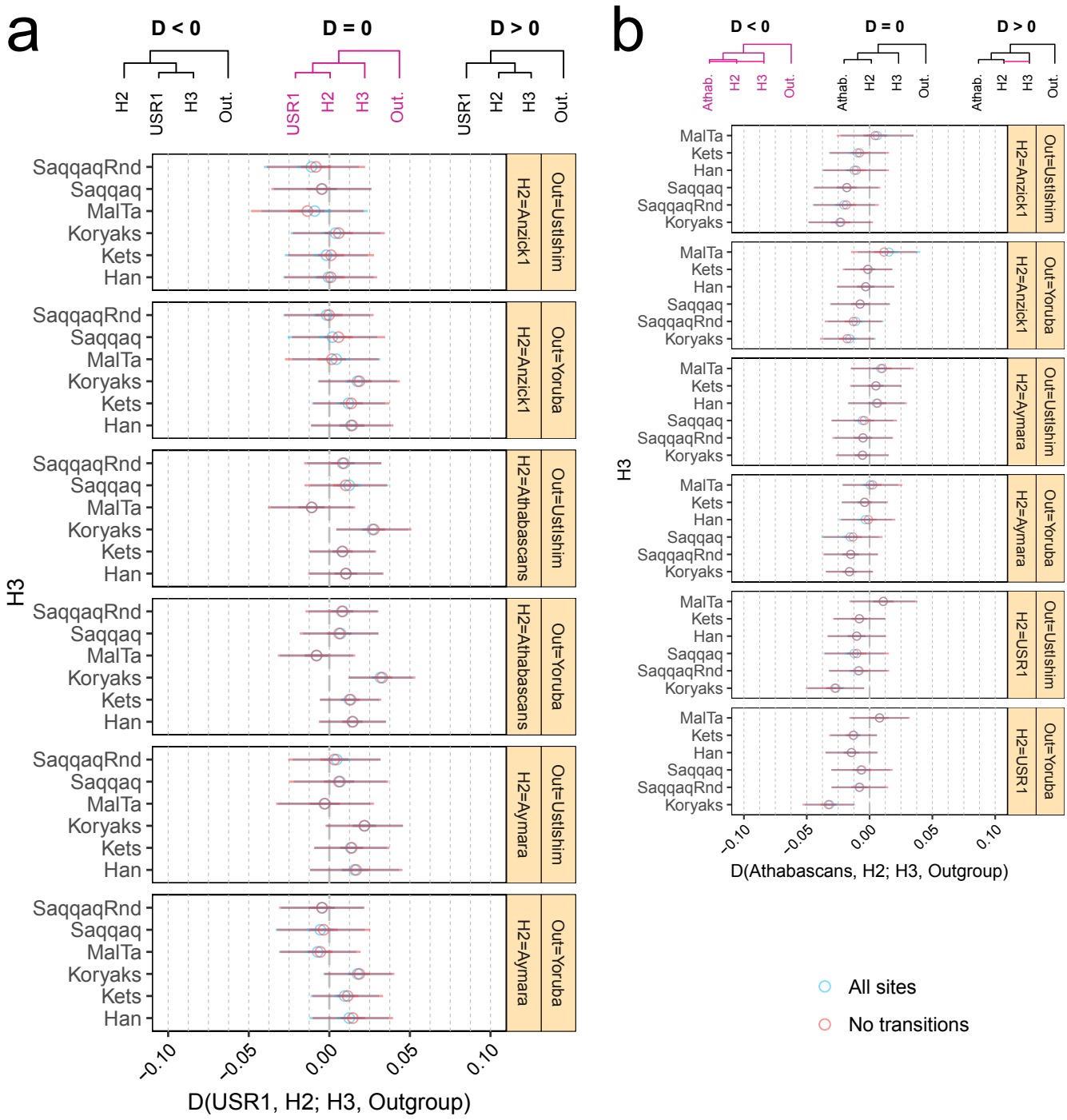

a
D < 0
D = 0
D > 0
H2
USR1
H3
Out.
USR1
H2
H3
Out.
USR1
H2
H3
Out.
SaqqaqRnd
Saqqaq
MalTa
Koryaks
Kets
Han
H2=Anzick1
Out=Ustlshim
H2=Anzick1
Out=Yoruba
H2=Athabascans
Out=Ustlshim
H2=Athabascans
Out=Yoruba
H2=Aymara
Out=Ustlshim
H2=Aymara
Out=Yoruba
H3
−0.10
−0.05
0.00
0.05
0.10
D(USR1, H2; H3, Outgroup)
b
D < 0
D = 0
D > 0
Athab.
H2
H3
Out.
MalTa
Kets
Han
Saqqaq
SaqqaqRnd
Koryaks
H2=Anzick1
Out=Ustlshim
H2=Anzick1
Out=Yoruba
H2=Aymara
Out=Ustlshim
H2=Aymara
Out=Yoruba
H2=USR1
Out=Ustlshim
H2=USR1
Out=Yoruba
H3
−0.10
−0.05
0.00
0.05
0.10
D(Athabascans, H2; H3, Outgroup)
All sites
No transitions

Figure S22. Pooled whole genome-based *D*-statistics testing the placement of Native Americans in the admixture graph model. a. We computed D(USR1, Native Americans; H3, Outgroup) in order to confirm that USR1 forms a clade with other Native Americans to the exclusion of Old World populations. **b.** We computed D(Athabascans, Native American; H3, Outgroup) in order to explore which was the best proxy for the Asian-related ancestry in Athabascans. Dendrograms above the plots represent the favored hypothesis depending on the value of *D*, and the expected under the admixture graph model is highlighted in pink. All tests were computed by including all sites (blue) and transversion polymorphisms only (red). Points represent *D* statistics and error bars represent one and ~3.3 standard errors (which corresponds to a *p*-value of *~0.001*). Each Saqqaq alternative version consists of one individual and each 'pooled population' contains the following number of individuals: Yoruba (n=1), Ust'ishim (n=1), Mal'ta (n=1), Han (n=1), Ket (n=2), Koryak (n=2), USR1 (n=1), Anzick1 (n=1), Aymara (n=1) and Athabascan (n=2).

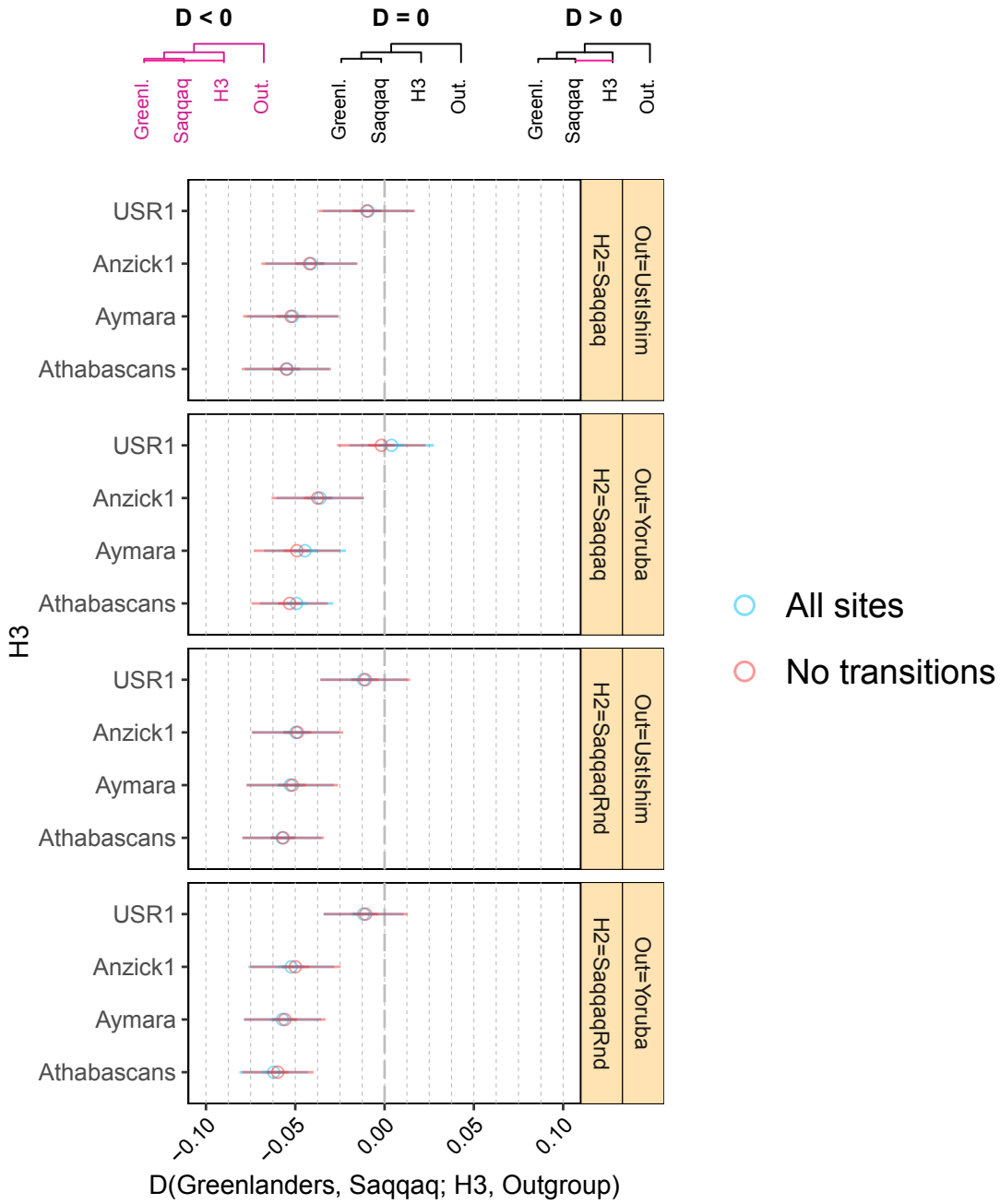

D < 0
D = 0
D > 0
Greenl.
Saqqaq
H3
Out.
USR1
Anzick1
Aymara
Athabascans
H2=Saqqaq
H2=SaqqaqRnd
Out=Ustlshim
Out=Yoruba
H3
−0.10
−0.05
0.00
0.05
0.10
D(Greenlanders, Saqqaq; H3, Outgroup)
All sites
No transitions

Figure S23. Pooled whole genome-based *D*-statistics testing the placement of Greenlandic Inuits in the admixture graph model. We computed D(Greenlanders, Saqqaq; Native American, Outgroup) in order to confirm that Greenlandic Inuit carry Native American admixture that is absent in Saqqaq. Dendrograms above the plots represent the favored hypothesis depending on the value of *D*, and the expected under the admixture graph model is highlighted in pink. All tests were computed by including all sites (blue) and transversion polymorphisms only (red). Points represent *D* statistics and error bars represent one and ~3.3 standard errors (which corresponds to a *p*-value of *~0.001*). All groups consist of one individual except Athabascans that consist of two.

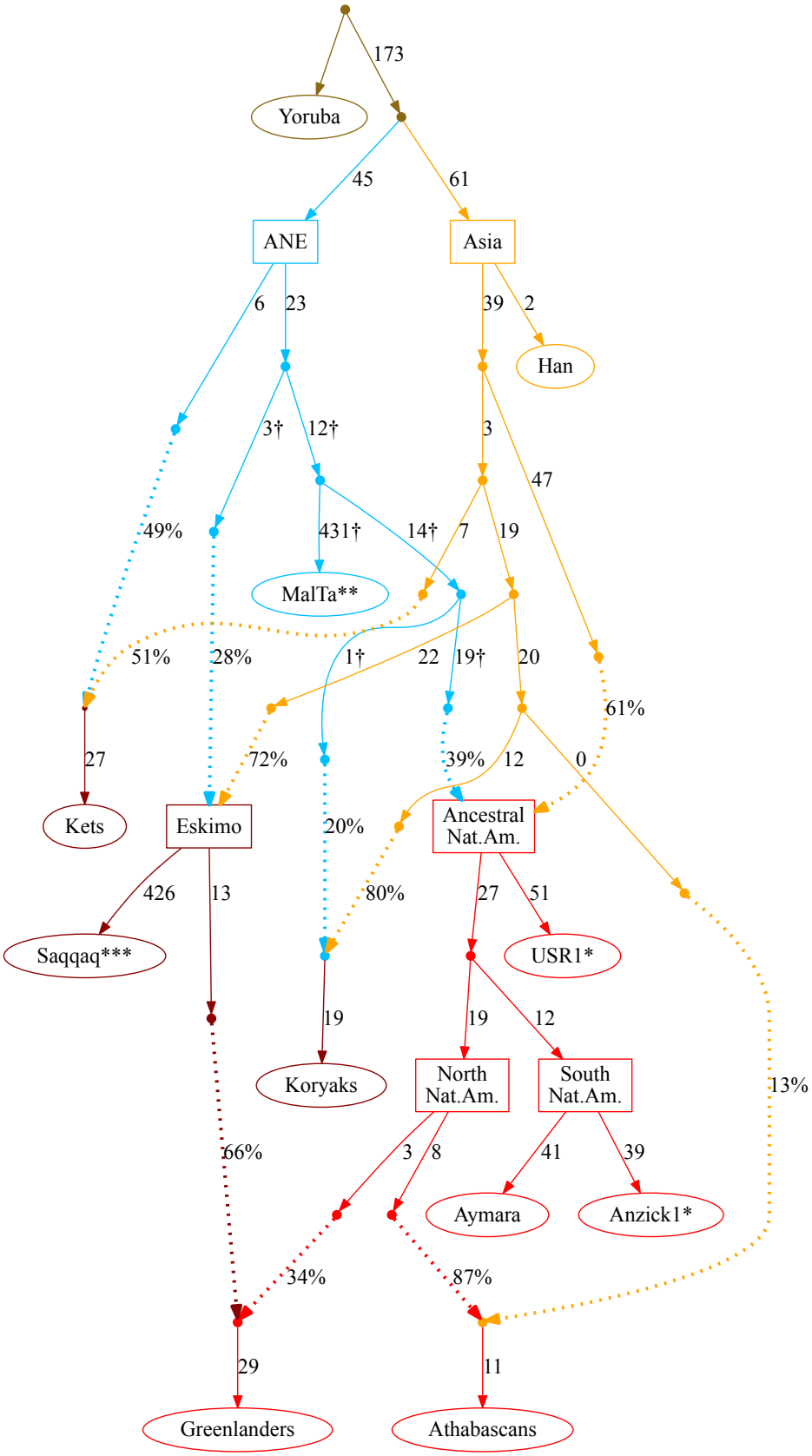

173
Yoruba
45
61
ANE
Asia
6
23
39
2
Han
3†
12†
3
47
49%
431†
14†
7
19
MalTa**
51%
28%
1†
22
19†
20
61%
27
72%
39%
12
0
Kets
Eskimo
20%
Ancestral
Nat.Am.
426
13
80%
27
51
Saqqaq***
USR1*
19
19
12
Koryaks
North
Nat.Am.
South
Nat.Am.
13%
66%
3
8
41
39
Aymara
Anzick1*
34%
87%
29
11
Greenlanders
Athabascans

Figure S24. A model for the formation of the different Native American populations using an alternative Saqqaq call set. We fitted an admixture graph by sequentially adding admixed leaves to a 'seed' graph including the Yoruba, Han, Mal'ta, Ket, USR1, Anzick1 and Aymara genomes. For each of the 'non-seed' admixed groups, we found the pair of edges that produced the best fitting graph, based on the fitting and maximum $|Z|$ scores (found to be 3.048 for the final graph). Box-shaped nodes represent metapopulations and sampled populations are shown in ellipse-shaped nodes respectively. * indicates that a population is represented by a single high-depth ancient genome, ** indicates that a population is represented by a single low-depth ancient genome and *** indicates that a random allele was sampled instead of using the Saqqaq called genotypes. † shows a subgraph whose structure we were not able to resolve with confidence due to additional sequencing and genotyping error in the Saqqaq genome (SI Section 17, Figure 3). The blue and orange subgraphs represent ancient north Eurasian and Asian ancestries, respectively.

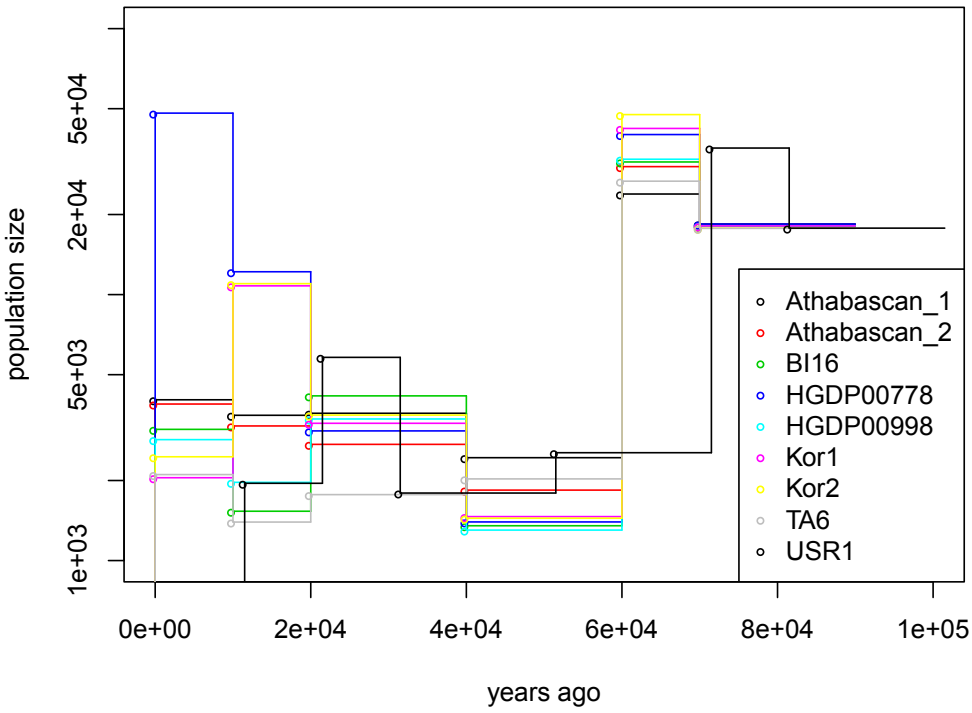

population size
1e+03
5e+03
2e+04
5e+04
0e+00
2e+04
4e+04
6e+04
8e+04
1e+05
years ago
Athabascan_1
Athabascan_2
BI16
HGDP00778
HGDP00998
Kor1
Kor2
TA6
USR1

Figure S25. Population size histories for a set of whole-genomes. Population sizes over six epochs with boundaries at 0 kya, 10 kya, 20 kya, 40 kya, 60 kya, 70 kya, ∞ were estimated using *diCal 2.0* with the aim of aiding in the pairwise comparisons (SI Section 18). Each line corresponds to one individual.

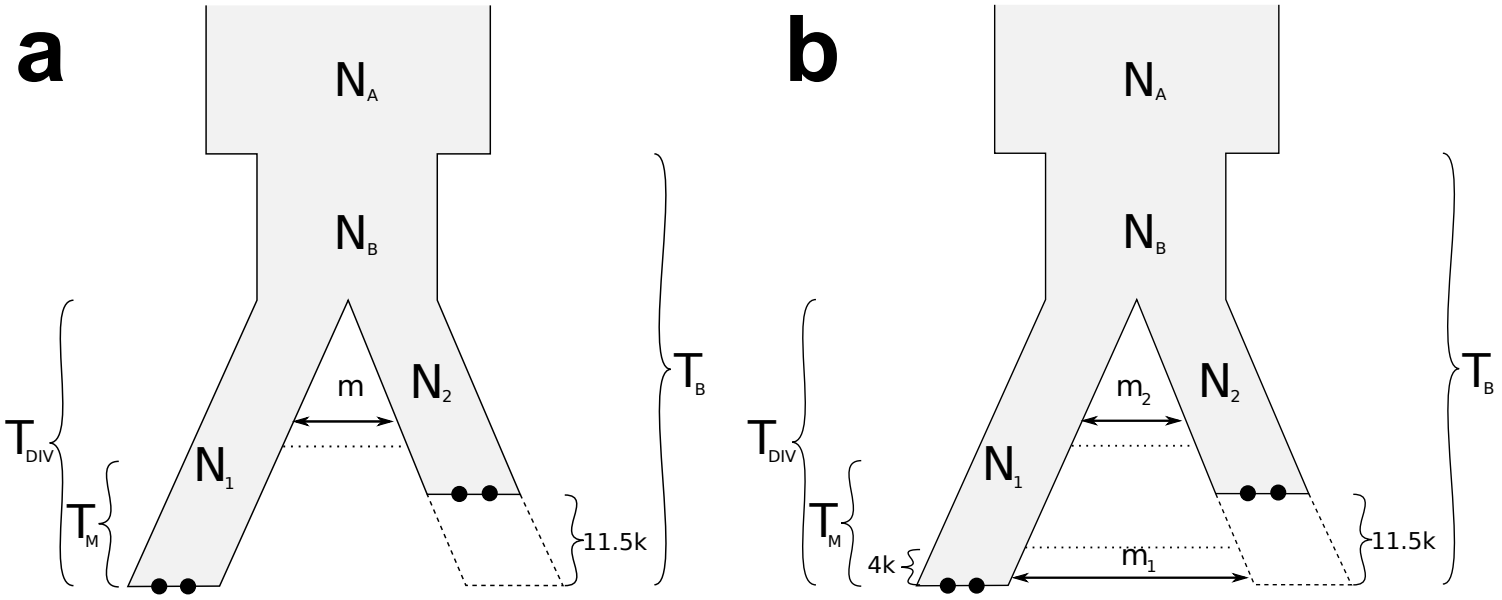

a
N_A
N_B
T_{DIV}
T_M
N_1
m
N_2
T_B
11.5k
b
N_A
N_B
T_{DIV}
T_M
N_1
m_2
N_2
T_B
4k
m_1
11.5k

Figure S26. Demographic models used for inference with diCal 2.0. a. a model for isolation starting at time T_{DIV}, with migration at rate m until time T_M . **b.** a second contact model. This is extension of the isolation with migration (IM) model from a., where we introduce a second period of gene flow at rate m_1 from 4 kya to the present. Black dots indicate the sampled haplotypes.

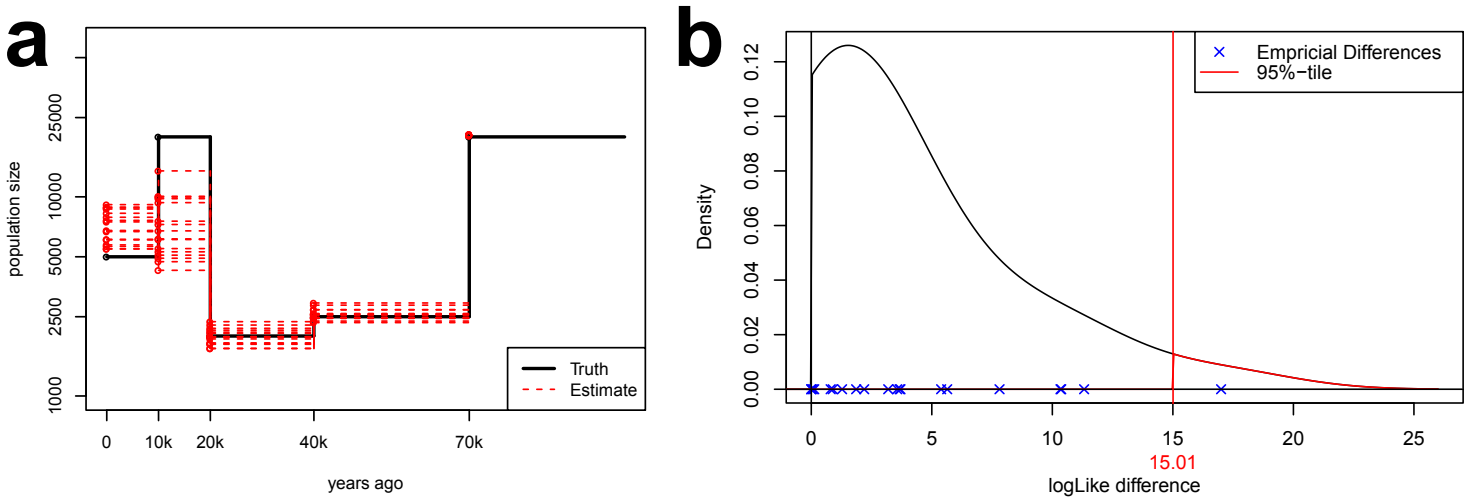

a
population size
1000
2500
5000
10000
25000
0
10k
20k
40k
70k
years ago
Truth
Estimate
b
Density
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0
5
10
15
20
25
15.01
logLike difference
Empricial Differences
95%−tile

Figure S27. Simulation results for validating population size estimates and hypothesis testing. a. *diCal 2.0* population size history estimates from simulated data. The true population size history is represented with a black line. The estimates from the 16 simulated datasets are represented with dashed red lines. **b.** the empirical distribution of the differences in log-likelihood from analyzing simulated data under a clean-split model and an isolation with migration model given in Tables S16 and S17. The data were simulated under a clean-split model.

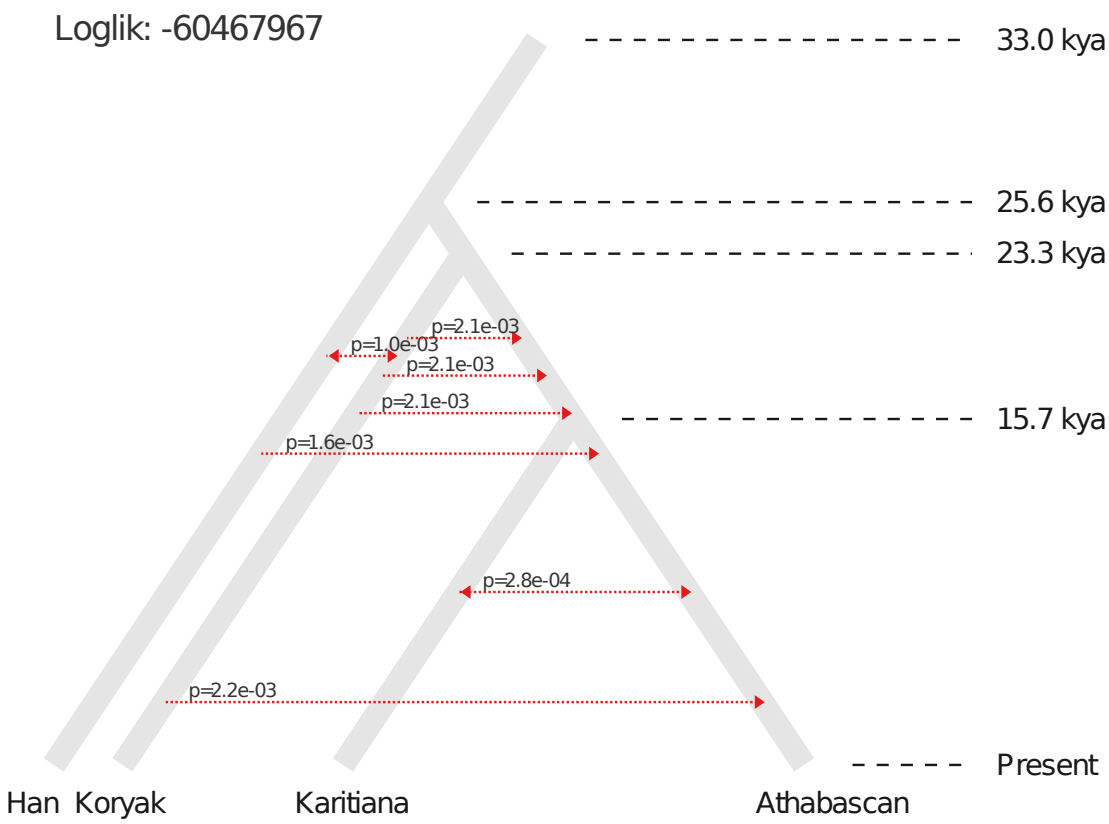

Loglik: -60467967
33.0 kya
25.6 kya
23.3 kya
p=2.1e-03
p=1.0e-03
p=2.1e-03
p=2.1e-03
15.7 kya
p=1.6e-03
p=2.8e-04
p=2.2e-03
Present
Han
Koryak
Karitiana
Athabascan

Figure S28. Inferred backbone demography for the SFS-based demographic inference strategy. We used *SMC++* and *momi2* to estimate the parameters of this model as detailed in SI Section 19. Note that migrations are modeled as consecutive pulses where a proportion of the lineages in one population are moved to another at a given generation. Pooled populations consist of the following numbers of individuals: Han (n=1), Koryaks (n=2), Karitiana (n=2), Athabascans (n=2).

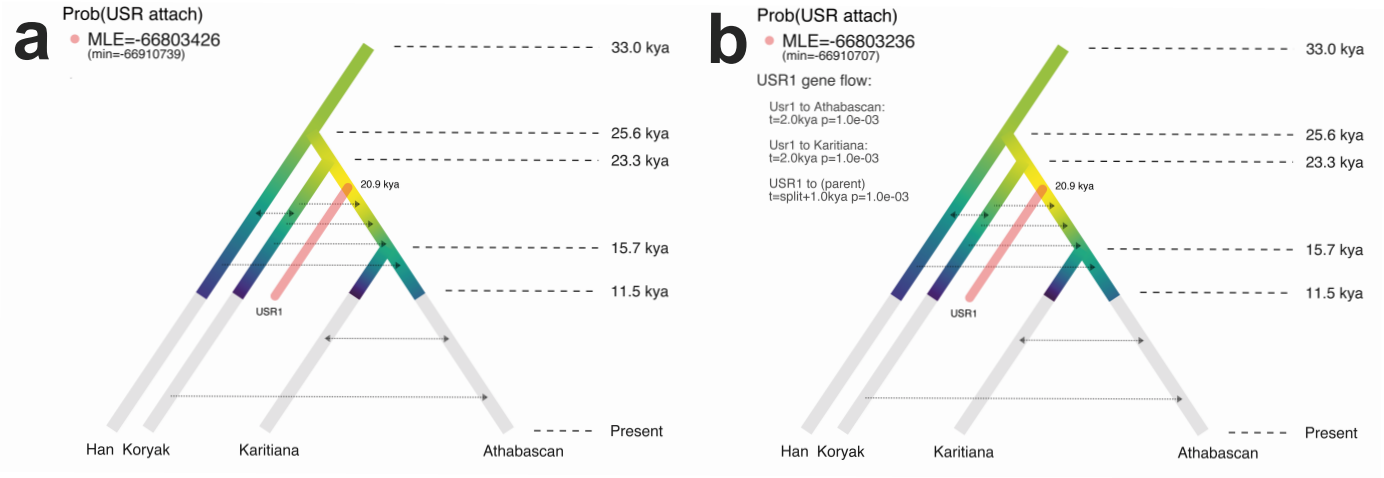

a
Prob(USR attach)
MLE=-66803426
(min=-66910739)
33.0 kya
25.6 kya
23.3 kya
20.9 kya
15.7 kya
11.5 kya
USR1
Present
Han
Koryak
Karitiana
Athabascan
b
Prob(USR attach)
MLE=-66803236
(min=-66910707)
USR1 gene flow:
Usr1 to Athabascan:
t=2.0kya p=1.0e-03
Usr1 to Karitiana:
t=2.0kya p=1.0e-03
USR1 to (parent)
t=split+1.0kya p=1.0e-03
33.0 kya
25.6 kya
23.3 kya
20.9 kya
15.7 kya
11.5 kya
USR1
Present
Han
Koryak
Karitiana
Athabascan

Figure S29. Likelihood heatmaps for the USR1 join-on point. The most likely join-on point for USR1 into the backbone demography (Figure S28) was found using an SFS-based maximum likelihood strategy (SI Section 19). Warmer colors represent a higher likelihood of USR1 joining at that point. **a.** Results from fitting USR1 as a clean split with no subsequent gene flow. **b.** Results from fitting USR1 with subsequent gene flow into the other three Native American lineages. Pooled populations consist of the following numbers of individuals: USR1 (n=1), Han (n=1), Koryaks (n=2), Karitiana (n=2), Athabascans (n=2).


Han
Koryak
Karitiana
Athabascan

Figure S30. Distribution of the likelihood surfaces for bootstrap analysis. 102 parametric bootstrap datesets were simulated and the whole inference process was repeated on each replicate. This figure shows the most likely join-on points for USR1, on all bootstrap replicates. Warmer colors represent a higher likelihood of USR1 joining at that given point.