HHS Public Access

40,000-Year-Old Individual from Asia Provides Insight into Early Population Structure in Eurasia

Melinda A. Yang[1,2], **Xing Gao**[1,2], **Christoph Theunert**[3,4], **Haowen Tong**[1], **Ayinuer Aximu-Petri**[2,4], **Birgit Nickel**[4], **Montgomery Slatkin**[3], **Matthias Meyer**[2,4], **Svante Pääbo**[2,4], **Janet Kelso**[2,4,*], and **Qiaomei Fu**[1,2,5,*]

[1]Key Laboratory of Vertebrate Evolution and Human Origins of Chinese Academy of Sciences, Institute of Vertebrate Paleontology and Paleoanthropology, Chinese Academy of Sciences, Beijing 100044, China

[2]Laboratory on Molecular Paleontology of the Max Planck Institute for Evolutionary Anthropology and the Institute of Vertebrate Paleontology and Paleoanthropology, Chinese Academy of Sciences, Beijing 100044, China

[3]Department of Integrative Biology, University of California Berkeley, Berkeley, Berkeley, CA 94720, USA

[4]Department of Evolutionary Genetics, Max Planck Institute for Evolutionary Anthropology, Leipzig 04103, Germany

[5]Lead Contact

SUMMARY

By at least 45,000 years before present, anatomically modern humans had spread across Eurasia [1–3], but it is not well known how diverse these early populations were and whether they contributed substantially to later people or represent early modern human expansions into Eurasia that left no surviving descendants today. Analyses of genome-wide data from several ancient individuals from Western Eurasia and Siberia have shown that some of these individuals have relationships to present-day Europeans [4, 5] while others did not contribute to present-day Eurasian populations [3,6]. As contributions from Upper Paleolithic populations in Eastern Eurasia to present-day humans and their relationship to other early Eurasians is not clear, we generated genome-wide data from a 40,000-year-old individual from Tianyuan Cave, China, [1, 7] to study

his relationship to ancient and present-day humans. We find that he is more related to present-day and ancient Asians than he is to Europeans, but he shares more alleles with a 35,000-year-old European individual than he shares with other ancient Europeans, indicating that the separation between early Europeans and early Asians was not a single population split. We also find that the Tianyuan individual shares more alleles with some Native American groups in South America than with Native Americans elsewhere, providing further support for population substructure in Asia [8] and suggesting that this persisted from 40,000 years ago until the colonization of the Americas. Our study of the Tianyuan individual highlights the complex migration and subdivision of early human populations in Eurasia.

In Brief

Yang et al. show that the genome of a 40 kya individual from Tianyuan cave near Beijing is more similar to Asians than to Europeans, past or present. His similarity to a 35 kya European and to individuals from some South American populations suggests a persistence of population structure in Asia that lasted until the colonization of the Americas.

Graphical Abstract
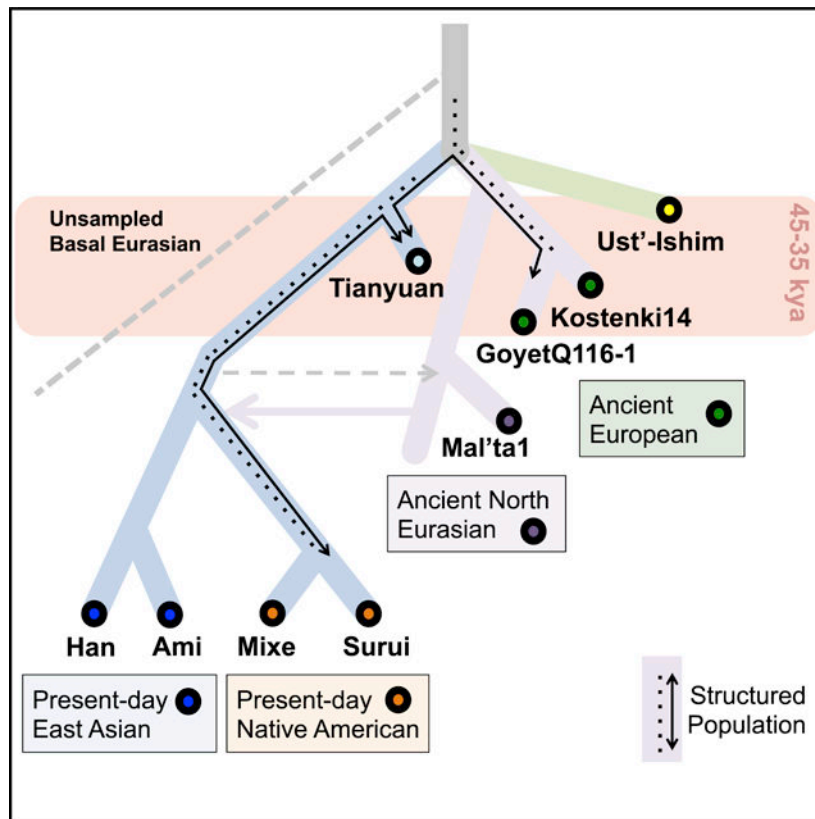
RESULTS AND DISCUSSION

Due to the high fraction of microbial sequences in DNA libraries generated from the Tianyuan individual [7], we used hybridization to oligonucleotide probes [6, 9] to enrich for

human DNA fragments carrying 3.7 million single-nucleotide polymorphisms (SNPs) (Tables S1A and S1B). We aligned captured sequences to the human reference genome *hg19,* and at the 2,228,374 sites covered by captured sequences, we obtained an average coverage of 2.98-fold (Tables S1B and S1C). We drew a random sequence for each site to represent the Tianyuan genome and merged this with published data from ancient (Table S1D) and present-day humans (Table S1E).

When sequencing ancient modern humans, it is important to account for contamination from present-day human DNA. We previously showed that DNA in these libraries has cytosine deamination patterns characteristic of ancient DNA and that human mtDNA contamination from these 15 libraries ranges from 0.2% to 1.5% [7]. To estimate nuclear contamination for the capture data, we used an approach [10] that takes advantage of the fact that the Tianyuan individual is male and thus carried one copy of the X chromosome. Assuming that polymorphic sites on the X chromosome are due to contamination rather than sequencing errors, we estimate nuclear contamination between 0.01% and 1.92% in the libraries used (Table S1A).

It has previously been shown using chromosome 21 that the Tianyuan individual shares more alleles with present-day Asians than with present-day Europeans [7]. We similarly find that the Tianyuan individual shares more alleles with present-day Eastern Eurasians, Oceanians, and Native Americans than with other present-day humans using the statistic [11] $f_3(Tianyuan, X; Mbuti),$ with the highest similarity to East and Southeast Asian populations (Figure 1A; Table S2A). However, present-day Europeans were found to carry a genetic component from a population that diverged from other non-Africans before they diverged from each other (a Basal Eurasian population [12]), such that analyses using only present-day Europeans may make the Tianyuan individual look more closely related to present-day Asians than he was. We therefore compared the Tianyuan individual to ancient Europeans who show no evidence of any Basal Eurasian ancestry [4] and are of an age similar to him (Kostenki14, GoyetQ116–1, and Vestonice16). The Tianyuan individual consistently shares more alleles with ancient and present-day East and Southeast Asians, as well as ancient and present-day Native Americans, than with either ancient or present-day Europeans (|Z| >3; Figure 2A; Tables S2B–S2D). We also find that ancient and present-day East and Southeast Asians and Native Americans are all more closely related to each other than they are to the Tianyuan individual (Figures 2B and S1; Tables S2B–S2D). Taken together, our results indicate that the Tianyuan individual is related to an ancestral group that contributed to all more recent populations with Asian ancestry. Also, the Tianyuan individual's age indicates that a genetic separation of Europe and Asia must have been earlier than 40,000 years ago. This is consistent with a split time of 40,000–80,000 years ago estimated for European and Asian populations based on mutation rates estimated from de novo mutations [13–15].

When combined with data from other early Upper Paleolithic individuals from Eurasia [3–6], our results show that several distinct populations existed in Eurasia before 35 kya. One population, represented by the 37,000-year-old Kostenki14, contributed genetic ancestry to present-day Europeans; a second population represented by the Tianyuan individual contributed to present-day East and Southeast Asians; and one or more additional

populations represented by the 45,000-year-old Ust'-Ishim and the 40,000-year-old Oase1 individuals did not contribute detectably to any present-day populations. Including the inferred Basal Eurasian population contributing to present-day Europeans and ancient Near East individuals [12,16], a minimum of four populations must therefore have coexisted in Eurasia before 35 kya.

Among the Upper Paleolithic individuals from Western Eurasia analyzed here, a 35,000-year-old individual from Belgium, GoyetQ116–1 [4], shares more alleles with the Tianyuan individual than any other Western Eurasian individual does $(f_3(Tianyuan, GoyetQ116–1; Mbuti) = 0.23$, Figure 1B and Table S2A; $D(GoyetQ116–1, Vestonice16; Tianyuan, Mbuti) > 0$, Figures 2C and S2A and Tables S3A and S3B). Furthermore, admixture between the populations to which GoyetQ116–1 and the Tianyuan individual belong improves the fit of a tree relating these populations to one another (Figure S1). The excess of alleles shared by the Tianyuan individual with GoyetQ116–1 compared to other ancient Europeans persists when we restrict the analysis to deaminated sequence fragments and when we exclude potentially deaminated bases (Figure S2Aii and S2Aiii; Table S3B), suggesting that DNA contamination from present-day humans and nucleotide misincorporation induced by cytosine deamination do not explain this allele sharing. We also find no evidence that the similarities between the Tianyuan individual and GoyetQ116–1 are due to sequencing error, data processing artifacts, or reference bias (Table S3G).

Notably, ancient European individuals related to GoyetQ116–1, such as the 19,000-year-old El Miron individual from Spain [4], do not share more alleles with the Tianyuan individual than other ancient European individuals do (Figure S2B). Present-day East and Southeast Asian populations do not share significantly more alleles with GoyetQ116–1 than they do with other ancient Western Eurasians (Figures 2C and S2A). GoyetQ116–1 carries a mitochondrial genome belonging to haplogroup M, and M-derived haplogroups can be found in present-day East Eurasian, Oceanian, and Native American populations but are almost completely absent in European populations [17,18]. These results suggest that despite the geographical distance between them, the Tianyuan individual and GoyetQ116–1 may share ancestry from a population that did not contribute ancestry to the other Upper Paleolithic Eurasians analyzed to date. Other younger connections between Eastern and Western Eurasia have also been found. Lipson and Reich [19] find that the 24,000-year-old Mal'ta1 [20] and 16,500-year-old AfontovaGora3 [4] from western Siberia and several 7,000- to 14,000-year-old Western Eurasian individuals show evidence of gene flow from a population related to the East and Southeast Asian Ami. We observe that the Eastern European hunter-gatherer Karelia [9], like the ancient Siberians and Western Eurasians, also show evidence of Asian gene flow. We also find that the pattern occurs for more East and Southeast Asian populations than just the Ami (Figures S2C and S2D; Tables S3C–S3F). Previous demographic inference studies [21–23] have inferred non-zero levels of migration between the ancestors of present-day European and Asian populations. Using the Tianyuan individual, we directly show that the separation of populations ancestral to more recent Europeans and Asians was a complex process that may have involved a sub-structured ancestral population and gene flow subsequent to geographic separation of populations.

Present-day Asian individuals carry ancestry from both Neanderthals and Denisovans [24–27]. We find that the Tianyuan individual carried about as much Neanderthal DNA as other Upper Paleolithic Eurasians (~4%–5%), which is more than that in present-day Eurasians (~1%–2%; Figures S3A and S3B; Table S4C) and is consistent with the hypothesis that purifying selection acting since introgression has reduced the amount of Neanderthal DNA in present-day genomes [4]. We do not detect Denisovan ancestry at the levels observed in Oceanian populations [26] in the genome of the Tianyuan individual (Tables S4A and S4B). However, due to insufficient power to detect low levels of Denisovan admixture, we cannot exclude that the Tianyuan individual carries DNA derived from Denisovans at levels similar to present-day mainland Asians [28] (Figure S3C). We also do not detect admixture from other putative archaic groups (Figures S3D–S3G), despite morphological suggestions of admixture from an archaic population [29].

Most Asian and Native American populations share similar numbers of alleles with the Tianyuan individual (Tables S2Bv and S2Dviii). However, three South American populations—the Surui and Karitiana in Brazil (''Amazonians'') and the Chane in northern Argentina and southern Bolivia—share more alleles with the Tianyuan individual than other Native American populations do (Figure 3A; Tables S2E, S2H, and S2J). The two Amazonian populations were recently shown to share more alleles with the present-day Papuan and Andamanese Onge than with other Native Americans [8, 30, 31] (Figures 3B and 3C), suggesting that at least two populations contributed ancestry to Native Americans in Central and South America. A 12,000-year-old individual from North America (Anzick-1 [32]) does not share more alleles with the Tianyuan individual (or with Oceanians or the Andamanese [8]) than with other Native Americans (Figures 3A–3C; Tables S2Dviii and S2J). The Surui and Chane show the highest levels of allele sharing with the Tianyuan individual *(D(Surui/ Chane, Mixe, Tianyuan, Mbuti)* = 0.02, Z > 3; Table S2J), which is higher than, or similar to, levels of allele sharing with the Papuan or Onge (Table S2J). Using an analysis robust to uncertainty of the exact population history [9], we find that the Amazonians can be described as a mixture of other Native American populations and 9%–15% of an ancestral population related to the Tianyuan individual, the Papuan, or the Onge (SE 4%–10%; Table S2G). Although the SE is high, we note that the Amazonians are consistently modeled as a mixture of other Native Americans and the Tianyuan individual, the Papuan, or the Onge. The mixture proportion estimates are also similar across all analyses, indicating that the relationship between the Tianyuan individual and the Amazonians is similar to that reported between the Papuan and the Amazonians and Onge and the Amazonians.

We also studied a model relating the Tianyuan individual to other ancient individuals and present-day populations using a base model including the Altai Neanderthal, Denisovan, Ust'-Ishim, and Kostenki14 from Mallick et al. [33]. We added the Tianyuan individual, Mal'ta1, and the present-day Ami, Mixe, Surui, and Papuan. Because present-day European populations have recent ancestry from an unknown Basal Eurasian population, we use Kostenki14, which has recently been shown to have no Basal Eurasian ancestry [4] to represent Europeans. We caution that our model (Figure 3D) is unlikely to reflect the true population relationships, as we cannot model many demographic features, such as population structure or continuous migration [11]. Intriguingly, however, within this simple

model, we find that all three South American populations can be modeled as sharing ancestry not only with other Native Americans, but also with populations related to the Tianyuan individual and the Papuan (Figure 3D; Table S2I). The fact that the Tianyuan individual, who lived in mainland Asia about 40,000 years ago, has affinities to some South American populations that is as strong as or stronger than that observed for the Papuan and Onge suggests that a population related to the Tianyuan individual, as well as to the present-day Papuan and Onge, was once widespread in eastern Asia. This group or another Asian population related to this group persisted at least until the colonization of the Americas and contributed to the genomes of some Native American populations.

We show that the Tianyuan individual is more closely related to ancient and present-day East and Southeast Asians than to either ancient or present-day Europeans. To test whether he is from a population that is directly ancestral to any present-day East or Southeast Asians, we modified the test of direct ancestry described in Rasmussen et al. [32] to account for low-coverage sequence data, contamination, and sequencing error. We find that that the Tianyuan individual is not from a population that is directly ancestral to any group of present-day East or Southeast Asians ($c > 0$, $c = 0$ rejected with $p < 0.00001$ for every comparison in Table 1), but rather belonged to a population that diverged from the population that contributed to present-day East and Southeast Asians. This is consistent with his unique ties to the 35,000-year-old GoyetQ116–1 and to some South American populations, which are not observed in present-day East and Southeast Asians.

The Tianyuan genome, as well as the genomes of individuals that lived at a similar time in Europe, shows that Asian and European populations had already begun to diverge at least by 40,000 years ago. At that time, modern human populations in Eurasia were subdivided not only into populations ancestral to present-day Asians and Europeans, but also into populations that did not contribute detectably to present-day populations. The shared ancestry between the Tianyuan individual and a 35,000-year-old individual in Europe and present-day Native American groups in South America further highlights that both substructure and population contacts have characterized the population history in Eurasia.

STAR★METHODS

CONTACT FOR REAGENT AND RESOURCE SHARING

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Qiaomei Fu (fuqiaomei@ivpp.ac.cn).

METHOD DETAILS

Library Preparation

We selected 15 libraries (Table S1A) with the highest genomic coverage and lowest mitochondrial DNA (mtDNA) contamination estimates from 34 Tianyuan libraries previously generated [7] from the Tianyuan femur [1] (TY1301). These libraries are all doublestranded with a 4 base pair (bp) ''clean-room key'' sequence (5′-GTCT-3′) on both

adapters (Fu et al. [7]). Libraries were treated with uracil-DNA-glycosylase (UDG) and endonuclease (EndoVIII) as described under ''ds UDG'' in SI2 of Fu et al. [4].

In-solution capture of nuclear DNA

Since the content of human DNA in these libraries is only between 0.013% and 0.031% (Table S1A), we hybridized the libraries to four previously described panels of oligonucleotide probes [6] to enrich for 71,044,176 fragments (35 bp) carrying a total of approximately 3.7 million SNPs (Table S1C) Hybridization capture was performed separately for each panel. Enriched libraries were sequenced on the HiSeq2500 platform using a double index configuration (2×76bp). We trimmed identifying sequences and any trailing adapters, and merged forward and reverse sequences into fragments with a minimum overlap of 11 base pairs. Sequences were aligned to the human reference genome (hg19) using *bwa* [36] and the following parameters: -n 0.01 and–l 16500. We removed duplicate fragments with the same start and end positions, keeping the fragment with the highest average base quality (Table S1B).

DNA contamination and preparation for analysis

MtDNA contamination estimates ranged from 0.2% to 1.5% [7]. We generated a second contamination estimate using a test for polymorphic sites on the X chromosome [39] since the Tianyuan individual (TY) is male [1, 7, 29]. For each library we identified 100 X chromosome SNPs covered by at least two reads on polymorphic sites. Factors such as sequence and alignment errors can lead to incorrect identification of variants [39], so we use ANGSD [10] to generate a contamination estimate using information from the same reads to calculate local alignment and sequence error rates. To determine the error rates for these genomic regions and the specific reads used for the analysis, we extracted 5 bp adjacent to each of the defined SNPs and calculated for each the fraction of nucleotides deviating from the reference genome and estimated the total error to range from 0.04%–0.52%, which we incorporated into the probabilistic model provided within ANGSD.

We scored alleles by randomly sampling one fragment covering each site containing a SNP, so long as the base quality was 20, and ignoring the first and last two bases of each fragment. Of the four SNP panels in '3.7M' (Table S1C), most of our analyses use the first three panels ('2.2M'), with one analysis using Panel 4. How Tianyuan reads were directly used in the test of direct ancestry is described further below.

Dataset

We used a subset of the Simons Genome Diversity Panel (SGDP) [33], focusing on 32 East and Southeast Asian (EAS) and Native American (AMER) samples (Table S1E) for most analyses, and using other present-day populations as needed. We also use the 1000 Genomes Project [34] in the test of direct ancestry, and the chimpanzee reference genome (*panTro2*) [35]. We used 31 ancient modern humans and two archaic humans with coverage greater than 0.05, as listed in Table S1D. Alleles were scored as described above, or genotyped as described elsewhere [4]. Data for individuals <5 kya were compiled here, all other ancient data was already compiled in [4]. We carried out analyses using either all sites or only transversions, with differences noted below.

QUANTIFICATION AND STATISTICAL ANALYSIS

f-statistics

All present-day and ancient humans were analyzed using ADMIXTOOLS [11], particularly *qp3Pop* (f_3-statistic) and *qpDstat* (D-statistic), with standard errors (s.e.) computed using a block jackknife (significance cut-off of $|Z|>3$ where Z = estimated value divided by the s.e.) [40]. For *qpDstat,* we use the default option "f4mode: NO."

We use the "outgroup" f_3-statistic, *f_3(Source1, Source2; Target),* to measure shared genetic drift between two source populations compared to an "outgroup," the target population [20]. Larger values indicate greater shared genetic drift between two source populations, which may reflect a closer relationship between these two sources. We used *f_3(TY, X; Mbuti),* where X refers to various ancient individuals and present-day populations (Table S2A; Figure 1).

In the D-statistic, *D(P1, P2; P3, Outgroup), D>0* indicates that P1 and P3 share more alleles than P2 and P3 share, and *D <0* indicates that P2 and P3 share more alleles than P1 and P3 share. Failing to reject *D = 0* ($|Z|<3$) indicates that P1 and P2 share a similar number of alleles with P3. A single D-statistic analysis can only compare the relative number of shared alleles between P1 and P3, and P2 and P3, but not between P1 and P2. Hence, a test for 'treeness' requires placing each population in turn into the position of P3 to look for consistent patterns of shared alleles between two populations.

We used D-statistics related to the Tianyuan individual as follows:

Comparisons to ancient Eurasians (EUR) and East Asians (EAS)—We find *D(EUR, EAS; TY, Mbuti)<0, D(TY, EUR; EAS, Mbuti)>0* and *D(TY, EAS; EUR, Mbuti) 0* (Tables S2Bi–S3Biii). *D(TY, EAS; EUR, Mbuti)<0* (Table S2Biii) mainly occurs for more recent EUR (and present-day Europeans), consistent with more shared ancestry between present-day Asians and Europeans discussed here and in [4,19]. Ancient Asian individuals (ANCEAS = DevilsGate1, Cho-khopani1/Mebrak/Samdzong; Table S1D) show similar patterns (Tables S2Civ–S2Cvi). Results for transversions are similar to that for all sites, except when underlined (Table S2Biii), but $|Z|>2$. The Tianyuan individual shares more alleles with Asian populations and individuals relative to ancient Eurasians of a comparable age (> 29 kya; Table S1D), confirming that he is genetically most similar to Asian populations.

Comparisons within Asians—We find *D(TY, EAS; EAS, Mbuti)<0* and *D(EAS, EAS; TY, Mbuti)~0* (Tables S2Biv and S2Bv). Deviations for the Uygur, Cambodians and Tu are consistent with recent gene flow [11, 37,41]. We find similar results for the ANCEAS (Tables S2Ci–S2Ciii), suggesting that they and the EAS form a clade with respect to the Tianyuan individual. Results are similar for all sites and for transversions only.

Comparisons to Native Americans (AMER, ancient Native American is Anzick-1)—We find *D(TY, AMER; EAS, Mbuti)<0, D(TY, EAS; AMER, Mbuti)<0* and *D(EAS, AMER; TY, Mbuti)   0* (Tables S2Di-S2Diii), suggesting that EAS and AMER

form a clade with each other with respect to the Tianyuan individual, but EAS share more with him than AMER share, consistent with AMER having a second ancestral component that is an outgroup to Asian populations [20]. Results are similar when using transversions only, except for Anzick-1, where the Z-score for *D(EAS, Anzick-1; TY, Mbuti)* changes from 3.2–3.4 to 1.7 (Table S2Diii). It is difficult to determine whether the loss of significance is due to deamination-induced errors or loss of statistical power. Different results for the Surui and Chane are consistent with later analyses.

We also find *D(EUR, AMER; TY, Mbuti)<0, D(TY, EUR; AMER, Mbuti) ≈ 0* for all but Mal'ta1 and Karelia where *D ≈ 0,* and *D(TY, AMER; EUR, Mbuti) ≈ 0* (Tables S2Div–S2Dvi). These results are consistent with AMER possessing ancient North Eurasian ancestry, represented by Mal'ta1 [20] and also found in the Karelia individual [16] – thus sharing a connection to individuals of European ancestry. Results for transversions are mostly consistent (|Z|>2.6), except in *D(TY, AMER; EUR, Mbuti),* where results for Kostenki14, Vestonice16, ElMiron and Villabruna are not significant (|Z|>1.6, Table S2Dvi).

Comparisons within Native Americans—Under a model where the ancestors of all AMER entered the Americas in a single wave, we would expect them to be equally distant from the Tianyuan individual. We find *D(TY, AMER; AMER, Mbuti)<0* and *D(AMER, AMER; TY, Mbuti)~0* (Tables S2Dvii–S2Dviii) except for the Mixtec, who may have more West Eurasian gene flow [8], and the Surui and Chane, suggesting a connection to the Tianyuan individual relative to the Mixe and Quechua (who may also have West Eurasian gene flow [8]). For transversions, results are similar (|Z|>2.8), except *D(Anzick-1, Mixe; Tianyuan, Mbuti)>0* (Z = 3.1), but we do not observe a similar pattern elsewhere.

Comparisons to the Onge and Papuan—The Onge acts similarly to EAS when compared to EUR, but shares a weaker connection to the Tianyuan individual than EAS (Table S2F). Results for the Papuan are biased when they are not in the P3 position because of Denisovan gene flow into Papuans of 4%–6% [26,27], but *D(TY, EUR; Papuan, Mbuti)>0* (Z1, Table S2F), suggesting the Papuan and Tianyuan individual share a connection relative to other ancient Eurasians.

Comparisons further exploring the South American connection—As in [8], we used *D(Amazonians = Surui/Karitiana, Central Americans; P3, Chimp),* with many different human populations as P3, and listed the highest D-statistics in Table S2E. The Tianyuan individual shares more alleles with Amazonians than with Central Americans, though we fail to reject *D(Karitiana, Mixe; Tianyuan, Mbuti/Chimp) = 0* (Table S2E). The Papuans consistently share more alleles with Amazonians than with Central Americans, but the Onge do not (Table S2E), which differs from a previous study [8], perhaps because we use a different dataset with fewer Onge individuals. When using only transversions, there is a loss of significance in the D-statistics shown in Table S2E, but the Tianyuan individual and Papuan still rank among the top ten highest D-statistics. Amazonians share a similarly strong connection to both the Tianyuan individual and Papuan [8].

We expand on the above results to test *D(AMER, AMER; P3, Mbuti/Chimp).* For P3, we focus on the Tianyuan individual and the Papuan and Onge, but also consider EAS and the

Chokhopani1 individual (Table S2H). Some EAS/Papuan/Onge share more alleles with the Surui and Karitiana than with the Quechua, but the Quechua samples are shown to be potentially admixed [8], though unlike the Mixtec, there is no consistent results in the D-statistic analyses here to indicate it is different relative to other Native Americans. Relative to the Mixe, the Surui share a connection to the Tianyuan individual and the Papuan, and they do not show a connection to the EAS (Table S2H).

Comparisons exploring connection to GoyetQ116–1—Comparing ancient Eurasians of a comparable age to the Tianyuan individual (ANCEUR = Ust'-Ishim, Oase1, Kostenki14, GoyetQ116–1) to younger ancient Eurasians (X), we find *D(TY, X; ANCEUR, Mbuti), D(TY, ANCEUR; X, Mbuti)* and *D(ANCEUR, X; TY, Mbuti)* largely shows expected relationships depending on whether the individual possesses European or greater archaic (Oase1 [6]) ancestry, except *D(GoyetQ116–1, X; TY, Mbuti)* is often positive (Table S3A, 'Z3').

To check for robustness of the GoyetQ116–1 result, we test for shared errors in GoyetQ116–1 and the Tianyuan individual related to ancient DNA (aDNA) damage, DNA contamination, or data processing, rather than past demographic processes. We find similar results when we restrict to only transversions ($|Z|>2.2$, D value remains similar; Table S3Bii), to avoid sites where differences might be errors from cytosine deamination due to ancient DNA damage rather than an actual allele from the ancient individual.

Restricting to damaged fragments reduces possible present-day DNA contamination, as it only merges sequenced DNA fragments showing characteristics of aDNA damage. Tianyuan libraries were double-stranded and uracil-DNA-glycosylase ('ds UDG') treated, meaning damage restriction could not be performed, but GoyetQ116–1 libraries were either single-stranded with UDG ('ss UDG') treatment or double-stranded without UDG ('ds noUDG') treatment, so damage restriction could be performed (Box 2.1 in SI2 of Fu et al. [4]). The Tianyuan/GoyetQ116–1 connection is still found using damage-restricted fragments (Table S3Biii; Figure S2Aiii). We also find *D(Stuttgart, French/Sardinian; TY, Mbuti)~0* (Z = 2.8/0.8) and we observe a weaker relationship between EAS and GoyetQ116–1 than between the Tianyuan individual and GoyetQ116–1, relative to other ancient West Eurasians (Figures 2C and S2A), so results are not consistent with present-day contamination.

All ancient individuals were similarly processed, but only GoyetQ116–1 shows the Tianyuan connection, making it unlikely the signal is an artifact of data processing. Our results were robust to outgroup choice except when using archaic individuals (Tables S3Bi–S3Biii). However, slight differences in how GoyetQ116–1 and the Tianyuan individual relate to these archaic individuals may mask the signal (Table S4A). Shared higher archaic admixture proportion in the Tianyuan individual and GoyetQ116–1 cannot explain the signal, as we do not observe that the Denisovan or Altai Neanderthal share more alleles with the Tianyuan individual and GoyetQ116–1 than with other ancient Eurasians of a comparable age (Tables S4A and S4B). No correlations in sequencing error rate, reference bias, or GC-content bias for GoyetQ116–1 or the Tianyuan individual can explain why the connection is only observed for these two individuals (Table S3G).

D(GoyetQ116–1, Kostenki14; TY, EAS) 0, and *D(GoyetQ116–1, Vestonice16; TY, EAS)~0* (Tables S3Ci and S3Cii). Thus, it is unclear whether the connection to GoyetQ116–1 is unique to the Tianyuan individual or also found in EAS but weaker due to dilution of the ancestral population of EAS. While EAS may show a weak connection to GoyetQ116–1, *D(GoyetQ116–1, Vestonice16; AMER, Mbuti)* is much closer to zero (Figure 2C). It is difficult to determine what other relationships may be having an effect on EAS and AMER, as they have complex relationships with more recent individuals and populations with European ties [4, 19, 20]. Comparing to the individual with the closest relationship to GoyetQ116–1, ElMiron [4], we find *D(ElMiron, EUR; TY, Mbuti)~0* (Figure S2B), suggesting the connection does not extend to her.

Comparisons to archaic humans (ARC = Altai Neanderthal/Denisovan)—We find *D(TY, Mbuti, ARC, Chimp) = 0.02/–0.001 (Z = 4.1/–0.2)*, similar to other Eurasians [25, 26], and *D(TY, non-Africans, ARC, Chimp)~0* (Table S4A), except for Oase1, an individual with higher Neanderthal ancestry due to admixture in his recent past [6] and the Papuan, who possesses Denisovan ancestry [26, 27]. GoyetQ116–1 shares more alleles with the Denisovan than ElMiron, Villabruna, Stuttgart and the present-day Sardinians share, and more alleles with the Altai Neanderthal than Satsurblia, Stuttgart, and present-day French and Sardinians share, but these signals disappear for transversions. One possibility is that false positives arise from deamination-induced errors in GoyetQ116–1 and the Denisovan, though we cannot rule out loss of statistical power from the reduced number of transversion SNPs. There is no evidence in the f_4-ratio test (Table S4C; Figures S3A and S3B) or in previous studies [4] to suggest GoyetQ116–1 has more archaic ancestry than other Upper Paleolithic Eurasians. To increase the power to detect archaic admixture we used Panel 4 of the '3.7M' dataset, which are informative about archaic ancestry, but we find no notable differences between ancient Eurasians (Table S4B).

D(TY, non-African; African, Chimp) tends to be negative for more recent non-Africans, but we also find that all more ancient non-African individuals share less alleles with African populations (including the ancient African individual Mota [42]) than more recent non-Africans share (Figure S3D). One explanation for this pattern is admixture into ancient Eurasians from a population more diverged than all humans presently sampled, so we investigate potential causes of this D-statistic signal. We find support that the ghost admixture observed is potentially representative of the difference in archaic admixture described between older and more recent Eurasians in [4]. The pattern is remarkably similar to the pattern found in the archaic admixture proportion (Figures S3A, S3B, and S3E). In a regression comparing the two statistics, we find they are highly correlated (r = 0.9, p < 0.001; Figure S3F).

Comparisons of Asians to more recent West and North Eurasians—Comparing ancient North (ANE = Mal'ta1, AfontovaGora3) and <14 kya West (EUR<14 kya = Villabruna, Bichon, Rochedane, Ran-chot88, Loschbour, LaBrana1, Hungarian.KO1, Karelia and Motala12) Eurasians to EAS and the Tianyuan individual, as well as other Eurasians (EUR = Kostenki14, Vestonice16, ElMiron, Villabruna), we find the ANE and EUR<14 kya form a clade with the EUR (Tables S3Di, S3Dii, S3Ei, and S3Eii), but *D(EUR,*

ANE/some EUR<14 kya; *EAS, Mbuti)<0* (Tables S3Diii and S3Eiii; Figures S2C and S2D), suggesting a connection between EAS and ANE, and EAS and some EUR<14 kya. We refer to the Bichon, Loschbour, LaBrana1, Hungarian.KO1 as the VWE, and Villabruna, Rochedane and Ranchot88 as the VNE in further analyses. The VNE do not tend to show a connection to EAS, whereas the VWE do, shown in Table S3Eiii and in [4]. The ANE, Motala12 and Karelia show a connection to the EAS, but a stronger connection to the AMER, while the VWE show a similar connection to both the EAS and AMER (Figures S2C and S2D). The Tianyuan individual may show a slight connection to the ANE (Table S3Diii; Figure S2C), but the single significant result, *D(Kostenki14, Mal'tal, TY, Mbuti)<0,* is not significant when using transversions only. No connection is observed for the VWE (Table S3Eiii; Figure S2D).

Lazaridis et al. [16] argue that West European hunter-gatherers (WHG: Loschbour, Hungarian.KO1, and LaBrana1) can be considered a mixture of populations related to East European hunter-gatherers (EHG: Karelia) and an older hunter-gatherer from Switzerland (Bichon, SI7 of Lazaridis et al. [16]). Here, we ask whether the Asian signal we observe in the VWE (composed of the WHG and Bichon) is related to the EHG's connection to EAS. First, [4] and we find that the Bichon individual also shares a connection to EAS. Second, for Karelia, we observe larger D values for AMER than for EAS, (Figure S2D; Table S3Eiii), likely related to the influence of ANE on Karelia [16, 20]. This difference is not observed in the WHG (Figure S2D). In contrast, Motala12 shows similar results to the WHG but does have elevated D values when compared to AMER (Figure S2D), suggesting Motala12 is connected to Karelia or the ANE. The Asian connection in the VWE cannot be explained by a connection to Karelia or the ANE.

Using *D(EUR<14* kya, *Kostenki14/GoyetQ116–1, TY, EAS)* (Table S3C), we show that EAS share more alleles with EUR<14 kya than the Tianyuan individual shares with them, relative to Kostenki14 and GoyetQ116–1, further supporting connections between EAS and ANE, the VWE, Karelia and Motala12. We note that EAS also share a connection to Satsurblia, Stuttgart, and the French and Sardinian, indicating other potential interactions unrelated to ANE or VWE.

We also used *qpF4ratio* to determine two f_4-ratios:

Determining archaic admixture proportion—Fu et al. [4] assumed that West and Central Africans (individuals from the Yoruba, Mbuti and Mende populations were grouped together as a single population) are outgroup to the East African Dinka population and present-day non-Africans. Then the Dinka represent a population with no archaic ancestry and Neanderthals and Denisovans ('Archaic') represent a population with complete archaic ancestry. Non-African individuals or populations (X) are then treated as a mixture of non-archaic human ancestry (F_{AFR}) and archaic ancestry (F_{ARC}). The constructed f_4-ratio (F_{AFR}) estimates the proportion of non-archaic human ancestry in present-day non-Africans, leaving one minus this quantity as F_{ARC}, the proportion of archaic ancestry in present-day non-Africans. F_{afr} measures the fraction of how many more shared alleles are between X and West and Central Africans than are between Dinka and West and Central Africans, relative

to the amount they each share with the archaic population including the Altai Neanderthal and Denisovan (merged set in analysis). F_{ARC}, or β(X) is represented by:

$$\beta(X) = F_{ARC}(X) = 1 - F_{AFR}(X) = 1 - \frac{f_4(X, \text{ Archaic; West and Central Africans, Chimp})}{f_4(\text{ Dinka, Archaic; West and Central Africans, Chimp })} .$$

We find β(TY)~0.05, similar to that seen in Eurasians of a similar age (excluding Oase1; Table S4C; Figures S3A and S3B). The connection between the Denisovan and GoyetQ116–1 in Table S4A is not observed here (Table S4C; Figures S3A and S3B). Results are similar using all sites and transversions only (Figures S3A and S3B), and are consistent with the conclusion of [4], where β(X) decreased over time–with higher levels in older Eurasians.

Test for Denisovan ancestry—We also used a test for low levels of Denisovan admixture (where $R_D>1$ may indicate some Denisovan gene flow, [28]), but we do not find $R_D>1$ for EAS as observed in [28] and find high variance and lack of consistency between using all individuals versus a single individual from a population (Figure S3C), likely because our sample sizes are too small for this analysis. Since there is only one Tianyuan individual, we do not have the power to test for low Denisovan ancestry as shown for EAS and AMER [28].

Inferring mixture proportions without a specific phylogeny: We determine whether two reference and one target populations *(Ref1, Ref2; Target)* have data consistent with being related by *N-1* streams of ancestry to a set of outgroup populations (using *qpWave* [11]) and if so, we estimate the proportion of ancestry in the population represented by the target that is composed of the populations represented by the references (using *qpAdm* [9]). *qpAdm* uses correlations in f-statistics to determine mixture proportions into a target population from a set of reference populations, presuming certain conditions are satisfied (rank1 > 0.05, 00_p > 0.05,01_pnest < 0.05; Table S3F for more details) [9], using a block jackknife as described for the f-statistics above to calculate the s.e. It requires a set of outgroup populations that contain enough genetic variation to be informative on the relationships between reference and target populations, but to have had no recent gene flow from reference or target populations.

To estimate extra Asian ancestry in South Americans, we tested *(Tianyuan/Han/Papuan/Onge, AMER; AMER)* using the outgroup set listed in Table S2G. With AMER as both reference and target, we ask whether the target AMER have any genetic signal that cannot be explained completely by the reference AMER, but instead requires a proportion of ancestry explained by the Tianyuan individual, the Papuan or the Onge. We find that the Karitiana, Surui and Piapoco are best described as a mixture of other AMER and the Tianyuan individual, Han, Papuan or Onge, with the least combinations found for the Han (Table S2G).

To assess the connection between East and North/West Eurasia, we tested four scenarios using the outgroups **Out** and **Out-T** as described in Table S3F:

Scenario 1: We used *(Han/TY, ANE; EAS)* to test gene flow into EAS from a population related to ANE. We find no evidence for Scenario 1, using the Han or the Tianyuan individual, and using Mal'ta1 (Table S3Fi) or AfontovaGora3.

Scenario 2: We used *(Vestonice16, EAS/TY; ANE)* to test gene flow from EAS into a population related to ANE. We find Scenario 2 does fit the data, using the Han or the Tianyuan individual, and using Mal'ta1 (Table S3Fii) or AfontovaGora3. Mixture proportions are high ($f\_2 = 0.3$–0.5; Table S3Fii), suggesting a stronger connection than indicated by the D-statistic analysis (Figure S2C), perhaps because the EAS and/or Vestonice16 are not the most appropriate proxy source populations.

Scenario 3: We used *(Han/TY, VNE/VWE; Ami)* to test gene flow into EAS (represented by the Ami) from a population related to the VWE. We find the Ami cannot be described as a mixture of a population related to the Han or the Tianyuan individual, and any VWE or VNE individual (Table S3Fiii). These results are consistent for other EAS.

Scenario 4: We used *(Ami/Han/TY, VNE; VWE)* to test gene flow into a population related to the VWE from a population related to EAS. This scenario is supported (Table S3Fiv) when using **Out-T**, but for **Out**, we cannot reject a model where all of the VWE's ancestry can be explained by the VNE (Table S3Fiv; pnest > 0.05).

For Scenarios 1 and 2, we also consider other outgroup sets, as shown in Table S3Fv. Here, and in Scenarios 3 and 4, we use the Ami to represent EAS, and though we show results for the Ami, we find similar results using other EAS. In Table S3Fv, we show that the results from our *qpAdm* analyses are robust to choice of outgroup set. For Scenario 1, we observe that for every outgroup set, a model of gene flow into the Ami from a population related to Mal'ta1 is not supported. For Scenario 2, we find that a model where Mal'ta1 is a mixture of populations represented by Vestonice16 and the Ami is possible for all outgroup sets. Our outgroup set choice does not alter our results.

Estimating a maximum likelihood tree: We used *Treemix* [37], which uses a Gaussian approximation of genetic drift and allele frequency data, to estimate a maximum likelihood tree including the Tianyuan individual. *Treemix* uses the variance-covariance matrix of allele frequencies between populations to determine the branch lengths ('Drift Parameter' in Figures 2B and S1) and allows addition of admixture events in its framework. We rooted the tree using the Central African Mbuti population as an outgroup, tested several block sizes (bs = 250, 500,1000,1500 and 2000) and admixture events (m ≤ 3 shown in Figure S1), and removed SNPs where an individual had missing data. We found similar results for all bs and show results for bs = 500. We minimized the residual fit, calculated by using the residual covariance between each pair of populations, divided by the average s.e. across all pairs, to determine which generated trees fit the data. For each tree, we also generated 1,000 bootstrap replicates and determined the uncertainty of each node (Figure S1). Results were similar for all sites and transversions. We analyzed three sets:

Set 1: Tianyuan, Ust'-Ishim, Kostenki14, GoyetQ116–1, Vestonice16 and four EAS (Han, Dai, Ami, Atayal). We do not include present-day Europeans, as they may have an ancestral component not represented in sampled populations and individuals [12].

Set 2: Set 1, AMER (Mixe, Quechua, Chane, Karitiana and Surui, Anzick-1), Mal'ta1.

Set 3: Set 2 and 3–1 kya individuals from Nepal (merged to one population 'AncTib', Figure S1C). We also consider the unmerged

individuals as separate tips in the tree.

In all sets, adding admixture improved the fit of the tree (Figure S1). The first admixture event was either gene flow from the Tianyuan individual into GoyetQ116–1 (10%, Figure S1Aiii) or gene flow from the Mal'ta1 individual into AMER. The Tianyuan individual always grouped most closely with EAS (Set 1, Table S1A), EAS and AMER (Set 2, Table S1B), and EAS, AMER and AncTib (Set 3, Table S1C), and each of these groups form distinct clades. In Set 1, EAS have a shared drift parameter since splitting from the Tianyuan individual of ~0.01. When the Tianyuan individual and EAS are grouped together, their shared drift parameter is an order of magnitude smaller, ~0.007, indicating that the Tianyuan individual diverges from EAS much closer to the node separating East Eurasians and ancient Europeans than any of the EAS, who diverge from each other much more recently. We cannot resolve the polytomy for Ust'-Ishim, Europeans and Asians, as the node for Ust'-Ishim has low bootstrap certainty (Figure S1), and in Set 1, the shared drift parameter for Eurasians after the Ust'-Ishim separates is only 0.0002–0.0010. In Set 2, while Mal'ta1 consistently shares an admixture event with AMER (Figures S1B and S1C), consistent with AMER sharing a relationship to ANE [20], the placement of Mal'ta1 with Europeans is supported for few admixture events, but not well supported when more admixture events are included (Figure S1B). In Set 3, after including the Mal'ta1 and AMER admixture event, the AncTib diverges from the ancestral population leading to both EAS and AMER (Figure S1C). This is true for transversions and all sites, using the merged AncTib set or keeping them separate. Figure S1B section iii is the same as Figure 2B.

Comparing archaic admixture to a connection to Africans: We examined the relationship between *D(non-African, non-African; African, Chimp)* and the archaic admixture proportion β(X), using a linear regression (Figure S3F). Using *ms* [38], we replicated simulated data 2,000 times for the model described in Figure S3G. We then asked how likely it is that given *D(X$_1$ orX$_2$, African1, Archaic, Outgroup)>0,* the following two conditions hold:

1.   We fail to reject *D(X$_1$, X$_2$, Archaic, Outgroup) = 0* (similar to Tables S4A and S4B).

2.   D(X$_1$, X$_2$, African1, Outgroup)<0 and D(X$_1$, X$_2$, African2, Outgroup)<0 (similar to Figure S3D).

We find that 750 replicates show *D(X$_1$ or X$_2$, African1, Archaic, Outgroup)>0,* and of those 105 satisfied the two conditions, so p = 0.14, suggesting it is not unusual to observe our seemingly contradictory results in Tables S4A and S4B and Figure S3D.

Testing a model including the Tianyuan individual: We used *qpGraph* (Admixture Graph) [11] to test models of the relationship between the Tianyuan individual and other Eurasian and Native American humans. Admixture Graph tests a proposed tree, or graph, for its fit to the data ($|Z|<3$), by checking the f_2, f_3 and f_4-statistics for all possible pairs, triples and quadruples of populations and individuals included in the graph. Graphs can include admixture, where a population is a mixture of two populations already represented. Solid branch labels give the estimated genetic drift in f_2-units of squared frequency difference in parts per thousand, dotted branch labels give the mixture proportions, and tips give the ancient individuals and present-day populations (Figure S4).

We begin with the base of Figure 3 of [33] that includes the Altai Neanderthal, Denisovan and Kostenki14. We add the Ust'-Ishim and Tianyuan individuals (Figure S4A), with the Mbuti as an African population without archaic admixture. We add extra admixture events from a population related to the Altai Neanderthal to account for higher archaic ancestry in early Upper Paleolithic individuals [4]. It is unlikely that three separate admixture events did occur, but this is the simplest method of allowing higher archaic ancestry in older individuals in the model, without which, models including present-day populations do not fit the data. We show Kostenki14 and the Tianyuan individual as more closely related to each other than to Ust'-Ishim, but we find no difference using models where either Kostenki14 or the Tianyuan individual are more closely related to Ust'-Ishim than each other and emphasize that the estimated branch length for M3-EUR0 is 0 (Figure S4A), indicating the three individuals form a polytomy.

We add Mal'ta1 and present-day humans (the Ami, Mixe, Surui and Papuan) in turn to the base graph (Figures S4B–S4L), exploring the fit of the data at each possible branching point for the added group, as well as every possible mixture of two ancestral components possible within the tree. As each is added one by one, a limitation of the method is that gene flow can only be tested in one direction. For instance, by adding Mal'ta1 then the Ami, we cannot test gene flow from a population related to the Ami into a population leading to Malta1, so we also tested adding the Ami, followed by Mal'ta1 (Figures S4E–S4G). We find that one or both must be included as a mixture of two ancestral components, but no model fits where neither Mal'ta1 nor the Ami are admixed. See Figures S4E–S4G for major patterns. We then add the Mixe, followed by the Surui (Figures S4I–S4L) to all possible models and find similar patterns where they must be modeled as a mixture of Asian and Mal'ta1-related ancestry, and the Surui can be described as possessing additional admixture from a population related to the Tianyuan individual. Adding the Papuan as an unadmixed population or a mixture of two populations already represented in the graph does not result in any model that fits the data.However, allowing a third admixture event into the Surui, such that the Surui is composed of three ancestral populations related to the Mixe, the Papuan and the Tianyuan individual, results in a model that fits the data, two of which are shown in Figures S4M and S4N. We find no difference in the fit of the model in the ordering of the three population mixtures, and in the main text display a model where a population that is a mixture of populations related to the Tianyuan individual and the Papuan then admixes with a population related to the Mixe (Figures S4N and 3D), as it is more geographically parsimonious than two separate admixture events from Asia (Figure S4M) into early AMER.

We also test adding GoyetQ116–1 or the Onge to the final model (Figure S4N), but could not find a graph that fits the data, suggesting complex demographic parameters outside the parameter space of AdmixtureGraph (i.e., population structure, continuous migration, etc.) may play a role. However, we note that the best-fitting graphs (Z = –3.5) always place GoyetQ116–1 as a mixture of a population related to Europeans and the Tianyuan individual. The best-fitting graphs for the Onge (Z = –3.4) always include them as a mixture of populations related to the Ami and Papuan. Future data that informs on their relationship to other Asian and European populations may help to develop a working population model that includes both GoyetQ116–1 and the Onge. Finally, we substituted each of the present-day Native American populations in place of the Mixe and Surui in Figure S4N to test whether these graphs would fit the data for different pairs of AMER. We looked for models that fit the data (|maxZ| < 3), have sufficient divergence between populations contributing to the AMER (we use the AMI0-SUPA1 branch, $d > 10$), and have added admixture into the P2 AMER (f_1 and $f_2 > 0.001$). AMER with a similar pattern to the Surui are the Chane and Karitiana (Table S2i). A model where some Southern American populations have gene flow from a population related to the Tianyuan individual can fit the data, but we caution that this is not the only model possible for AMER and other unexplored models may fit the data equally well. From these models, however, we find some patterns that seem consistent with other analyses in our study.

A method for testing for direct ancestry with low coverage data: We developed a test of whether an ancient individual is directly ancestral to a present-day population. Our test is based on one presented in SI17 of [32]. It estimates the probability of coalescence on the branch leading to the archaic population. If that probability is 0, then the archaic sample is from a population directly ancestral to the present-day population [32].

At each site, let A and a denote the two alleles where A is defined to be the allele on one present-day chromosome. Assuming the archaic genotype is known, only two possible genealogical histories are possible: coalescence before the most recent common ancestor (MRCA) or no coalescence (Figure S3H). With 3 ancestral lineages in the MRCA, the folded site frequency spectrum is characterized by k_1, the probability that one allele differs from the other two. Let c be the probability of a coalescence in the archaic population before the MRCA. For each site, the archaic genotype is aa, Aa or AA (numbered $i = 0, 1$ and 2). Let p_I be the probability of a site with configuration i. Then, following [32], we obtain

$$p_0 = \frac{(1+c)k_1}{3} \quad p_1 = \frac{2(1-c)k_1}{3} \quad p_2 = 1 + \frac{(c-3)k_1}{3} \quad \text{(Equation 1)}$$

We cannot use Equation 1 directly because the sequence coverage is too low to call heterozygotes with confidence and because there is sequencing error and contamination from present day humans. By modeling these processes, we can still estimate k_I and c.

With no sequencing error, reads are equivalent to each other, so the configuration has u reads carrying the same allele as the present-day sample ($u = 0, \ldots, d$). If q_u is the probability of configuration u,

$$q_u = \sum_{i=0}^{2} \Pr(u \mid i) p_i \quad \text{(Equation 2)}$$

Because the P_i are functions of c and k_1, the q_u are also functions of c and k_1.

To account for contamination let f be the frequency of A in the contaminating population. The probability that a read carrying a is replaced by a contaminating read carrying A is Kf, and the probability that a read carrying A is replaced by a contaminating read carrying a is $K(1 - f)$. Contamination causes the number of As to increase by k_+ with $Binomial(d - u, Kf)$, and to decrease by k_- with $Binomial(u, K(1 - f))$. Since K is small, accounting for increases or decreases by at most two is likely sufficient, such that the probability of observing configuration u after accounting for contamination is

$$r_u = \Pr(k_+ = 2 \mid u - 2) q_{u-2} + \Pr(k_+ = 1 \mid u - 1) q_{u-1} \qquad \text{(Equation}$$
$$+ (1 - \Pr(k_+ = 2 \mid u) - \Pr(k_+ = 1 \mid u) - \Pr(k_- = 2 \mid u) - \Pr(k_- = 1 \mid u)) q_u + \Pr$$
$$(k_- = 1 \mid u + 1) q_{u+1} + \Pr(k_- = 2 \mid u + 2) q_{u+2}.$$

3)

To account for sequencing error, we assume the probability of sequencing error differs for each read. Let the observed configuration of reads be $V = \{j_1, \ldots, j_d\}$ where $j_i = 0$ for a and $j_i = 1$ for A. Associated with each site is a vector $E_1 = \{e_1, \ldots, e_d\}$ (e_i = the probability of a single sequencing error in each read, computed from Phred scores) and a matrix with elements e_{ij}, the probability of a sequencing error both in read i and in read j, where we assume independence of errors, i.e., $e_{ij} = e_i e_j$.

For a given V with v As, let J_1 be the set of reads carrying an A and J_0 be the set of reads carrying an a. Let e_+ be the total probability that one sequencing error in a read with an a increases by one the number of reads apparently carrying an $A: e_+ = \sum_{j \in J_0} e_j$. Let e_- be the total probability that one error in a read with an A will decrease by one the number of reads apparently carrying an A, $e_- = \sum_{j \in J_1} e_j$.

To account for two errors, let J_{11} be the set of pairs of reads that both carry an A, J_{10} be the set where one carries an A and the other an a (in either order), and J_{00} be the set where both carry an a. Define $e_{--} = \sum_{\{j, j'\} \in J_{11}} e_{jj}$ as the total probability that two errors reduce the number of reads apparently carrying an A by two $e_{++} = \sum_{\{j, j'\} \in J_{00}} e_{jj}$ as the total probability that two errors increase the number of reads apparently carrying an A by two, and $e_{+-} = \sum_{\{j, j'\} \in J_{10}} e_{jj}$ as the total probability that the errors cancel and maintain the same number of reads apparently carrying an A. Finally, define

$e_0 = 1 - \sum_{j=1}^{d} e_j - \sum_{j=1}^{d} \sum_{j'=1}^{j-1} e_{jj}$ as the probability that there is no sequencing error at the site. Then

$$\Pr(V) = \left(e_0 + e_{+-}\right)\frac{r_v}{\binom{d}{v}} + \frac{r_{v-2}e_{++}}{\binom{d}{v-2}} + \frac{r_{v-1}e_+}{\binom{d}{v-1}} + \frac{r_{v+1}e_-}{\binom{d}{v+1}} + \frac{r_{v+2}e_{--}}{\binom{d}{v+2}} \quad \text{(Equation 4)}$$

Substituting Equation 1 through Equation 3 into Equation 4, we obtain an expression for the probability of the observed configuration V as a function of c and k-i, given the error probabilities, the contamination rate and the allele frequencies in the contaminating population. We calculate the overall probability of the data (the likelihood) by multiplying across sites and use standard optimization methods to obtain the maximum likelihood estimates to test the hypothesis that $c = 0$ by doing a likelihood ratio test.

To apply to the Tianyuan sample, we use the Yoruba population (n = 108) as the outgroup, identifying either all polymorphic sites (Analysis A) or only those with a frequency of 5%–95% (Analysis B), and the Han Chinese population (n = 103) as the contaminating population (samples from [34]) and used one chromosome from each of the listed present-day populations in Table 1, using the dataset from [33] (filters: mapping quality <30, base quality <20, ignoring the first and last two bases of Tianyuan reads and sites with >10 reads). We use $K = 0.02$ (contamination rate, largest in Table S1A is 0.192). Table 1 shows maximum likelihood estimates for c and k_1 for Analysis A. Analysis B (not shown) shows similar results, with slightly elevated k_1 values because we ascertain for more common alleles. The model is robust to higher sequencing error and contamination rates, as doubling the site-specific sequencing error rate did not affect our results and doubling the contamination rate increased parameter estimates, but did not change the relative ordering of individuals.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGMENTS

REFERENCES

1. Shang H, Tong H, Zhang S, Chen F, and Trinkaus E (2007). An early modern human from Tianyuan Cave, Zhoukoudian, China. Proc. Natl. Acad. Sci. USA 104, 6573–6578. [PubMed: 17416672]

2. Hublin J-J (2015). The modern human colonization of western Eurasia: when and where? Quat. Sci. Rev. 118, 194–210.

3. Fu Q, Li H, Moorjani P, Jay F, Slepchenko SM, Bondarev AA, Johnson PLF, Aximu-Petri A, Prüfer K, de Filippo C, et al. (2014). Genome sequence of a 45,000-year-old modern human from western Siberia. Nature 514, 445–449. [PubMed: 25341783]

4. Fu Q, Posth C, Hajdinjak M, Petr M, Mallick S, Fernandes D, Furtwängler A, Haak W, Meyer M, Mittnik A, et al. (2016). The genetic history of Ice Age Europe. Nature 534, 200–205. [PubMed: 27135931]

5. Seguin-Orlando A, Korneliussen TS, Sikora M, Malaspinas AS, Manica A, Moltke I, Albrechtsen A, Ko A, Margaryan A, Moiseyev V, et al. (2014). Paleogenomics. Genomic structure in Europeans dating back at least 36,200 years. Science 346, 1113–1118. [PubMed: 25378462]

6. Fu Q, Hajdinjak M, Moldovan OT, Constantin S, Mallick S, Skoglund P, Patterson N, Rohland N, Lazaridis I, Nickel B, et al. (2015). An early modern human from Romania with a recent Neanderthal ancestor. Nature 524, 216–219. [PubMed: 26098372]

7. Fu Q, Meyer M, Gao X, Stenzel U, Burbano HA, Kelso J, and Pääbo S (2013). DNA analysis of an early modern human from Tianyuan Cave, China. Proc. Natl. Acad. Sci. USA 110, 2223–2227. [PubMed: 23341637]

8. Skoglund P, Mallick S, Bortolini MC, Chennagiri N, Hünemeier T, Petzl-Erler ML, Salzano FM, Patterson N, and Reich D (2015). Genetic evidence for two founding populations of the Americas. Nature 525, 104–108. [PubMed: 26196601]

9. Haak W, Lazaridis I, Patterson N, Rohland N, Mallick S, Llamas B, Brandt G, Nordenfelt S, Harney E, Stewardson K, et al. (2015). Massive migration from the steppe was a source for Indo-European languages in Europe. Nature 522, 207–211. [PubMed: 25731166]

10. Korneliussen TS, Albrechtsen A, and Nielsen R (2014). ANGSD: analysis of next generation sequencing data. BMC Bioinformatics 15, 356. [PubMed: 25420514]

11. Patterson N, Moorjani P, Luo Y, Mallick S, Rohland N, Zhan Y, Genschoreck T, Webster T, and Reich D (2012). Ancient admixture in human history. Genetics 192, 1065–1093. [PubMed: 22960212]

12. Lazaridis I, Patterson N, Mittnik A, Renaud G, Mallick S, Kirsanow K, Sudmant PH, Schraiber JG, Castellano S, Lipson M, et al. (2014) . Ancient human genomes suggest three ancestral populations for present-day Europeans. Nature 513, 409–413. [PubMed: 25230663]

13. Roach JC, Glusman G, Smit AF, Huff CD, Hubley R, Shannon PT, Rowen L, Pant KP, Goodman N, Bamshad M, et al. (2010). Analysis of genetic inheritance in a family quartet by whole-genome sequencing. Science 328, 636–639. [PubMed: 20220176]

14. Scally A, and Durbin R (2012). Revising the human mutation rate: implications for understanding human evolution. Nat. Rev. Genet. 13,745–753. [PubMed: 22965354]

15. Scally A (2016). The mutation rate in human evolution and demographic inference. Curr. Opin. Genet. Dev. 41, 36–43. [PubMed: 27589081]

16. Lazaridis I, Nadel D, Rollefson G, Merrett DC, Rohland N, Mallick S, Fernandes D, Novak M, Gamarra B, Sirak K, et al. (2016). Genomic insights into the origin of farming in the ancient Near East. Nature 536, 419–424. [PubMed: 27459054]

17. Posth C, Renaud G, Mittnik A, Drucker DG, Rougier H, Cupillard C, Valentin F, Thevenet C, Furtwängler A, Wißing C, et al. (2016). Pleistocene Mitochondrial Genomes Suggest a Single Major Dispersal of Non-Africans and a Late Glacial Population Turnover in Europe. Curr. Biol. 26, 827–833. [PubMed: 26853362]

18. Kivisild T (2015). Maternal ancestry and population history from whole mitochondrial genomes. Investig. Genet. 6, 3.

19. Lipson M, and Reich D (2017). A working model of the deep relationships of diverse modern human genetic lineages outside of Africa. Mol. Biol. Evol. 34, 889–902. [PubMed: 28074030]

20. Raghavan M, Skoglund P, Graf KE, Metspalu M, Albrechtsen A, Moltke I, Rasmussen S, Stafford TW Jr., Orlando L, Metspalu E, et al. (2014). Upper Palaeolithic Siberian genome reveals dual ancestry of Native Americans. Nature 505, 87–91. [PubMed: 24256729]

21. Gravel S, Henn BM, Gutenkunst RN, Indap AR, Marth GT, Clark AG, Yu F, Gibbs RA, and 1000 Genomes Project, and Bustamante CD (2011). Demographic history and rare allele sharing among human populations. Proc. Natl. Acad. Sci. USA 108, 11983–11988. [PubMed: 21730125]

22. Laval G, Patin E, Barreiro LB, and Quintana-Murci L (2010). Formulating a historical and demographic model of recent human evolution based on resequencing data from noncoding regions. PLoS ONE 5, e10284.

23. Schiffels S, and Durbin R (2014). Inferring human population size and separation history from multiple genome sequences. Nat. Genet. 46, 919–925. [PubMed: 24952747]

24. Green RE, Krause J, Briggs AW, Maricic T, Stenzel U, Kircher M, Patterson N, Li H, Zhai W, Fritz MH-Y, et al. (2010).Adraft sequence of the Neandertal genome. Science 328, 710–722. [PubMed: 20448178]

25. Prüer K, Racimo F, Patterson N, Jay F, Sankararaman S, Sawyer S, Heinze A, Renaud G, Sudmant PH, de Filippo C, et al. (2014). The complete genome sequence of a Neanderthal from the Altai Mountains. Nature 505, 43–49. [PubMed: 24352235]

26. Meyer M, Kircher M, Gansauge M-T, Li H, Racimo F, Mallick S, Schraiber JG, Jay F, Prufer K, de Filippo C, et al. (2012). A high- coverage genome sequence from an archaic Denisovan individual. Science 338, 222–226. [PubMed: 22936568]

27. Reich D, Green RE, Kircher M, Krause J, Patterson N, Durand EY, Viola B, Briggs AW, Stenzel U, Johnson PLF, et al. (2010). Genetic history ofan archaic hominin groupfrom Denisova Cave in Siberia. Nature 468, 1053–1060. [PubMed: 21179161]

28. Qin P, and Stoneking M (2015). Denisovan Ancestry in East Eurasian and Native American Populations. Mol. Biol. Evol. 32, 2665–2674. [PubMed: 26104010]

29. Shang H, and Trinkaus E (2010). The Early Modern Human from Tianyuan Cave, China. In The Early Modern Human from Tianyuan Cave, China DG. Steele, ed. (Texas A&M University Press).

30. Skoglund P, and Reich D (2016). A genomic view ofthe peopling of the Americas. Curr. Opin. Genet. Dev. 41, 27–35. [PubMed: 27507099]

31. Raghavan M, Steinrücken M, Harris K, Schiffels S, Rasmussen S, DeGiorgio M, Albrechtsen A, Valdiosera C, Ávila-Arcos MC, Malaspinas A-S, et al. (2015). Genomic evidence for the Pleistocene and recent population history ofNativeAmericans. Science 349, aab3884.

32. Rasmussen M, Anzick SL, Waters MR, Skoglund P, DeGiorgio M, Stafford TW Jr., Rasmussen S, Moltke I, Albrechtsen A, Doyle SM, et al. (2014). The genome of a Late Pleistocene human from a Clovis burial site in western Montana. Nature 506, 225–229. [PubMed: 24522598]

33. Mallick S, Li H, Lipson M, Mathieson I, Gymrek M, Racimo F, Zhao M, Chennagiri N, Nordenfelt S, Tandon A, et al. (2016). The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. Nature 538, 201–206. [PubMed: 27654912]

34 1000. Genomes Project Consortium (2015). A global reference for human genetic variation. Nature 526, 68–74. [PubMed: 26432245]

35. ChimpanzeeSequencing andAnalysisConsortium (2005). Initial sequence of the chimpanzee genome and comparison with the human genome. Nature 437, 69–87. [PubMed: 16136131]

36. Li H, and Durbin R (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25, 1754–1760. [PubMed: 19451168]

37. Pickrell JK, and Pritchard JK (2012). Inferenceofpopulation splitsand mixtures from genome-wide allele frequency data. PLoS Genet. 8, e1002967.

38. Hudson RR (2002). Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics 18, 337–338. [PubMed: 11847089]

39. Rasmussen M, Guo X, Wang Y, Lohmueller KE, Rasmussen S, Albrechtsen A, Skotte L, Lindgreen S, Metspalu M, Jombart T, et al. (2011). An Aboriginal Australian genome reveals separate human dispersals into Asia. Science 334, 94–98. [PubMed: 21940856]

40. Busing FMTA, Meijer E, and Leeden RVD (1999). Delete-m jack- knife for unequal m. Stat. Comput. 9, 3–8.

41. Qin P, Zhou Y, Lou H, Lu D, Yang X, Wang Y, Jin L, Chung YJ, and Xu S (2015). Quantitating and dating recent gene flow between European and East Asian populations. Sci. Rep. 5, 9500. [PubMed: 25833680]

42. Gallego Llorente M, Jones ER, Eriksson A, Siska V, Arthur KW, Arthur JW, Curtis MC, Stock JT, Coltorti M, Pieruccini P, et al. (2015). Ancient Ethiopian genome reveals extensive Eurasian admixture throughout the African continent. Science 350, 820–822. [PubMed: 26449472]

Highlights

- By 40,000 years ago, 4 subpopulations of modern humans were established in Eurasia

- The Tianyuan individual is more related to Asians than to past and present Europeans

- He is not, however, equally similar to all early humans in Europe

- His genetic similarity to some South Americans suggests early Asian population structure
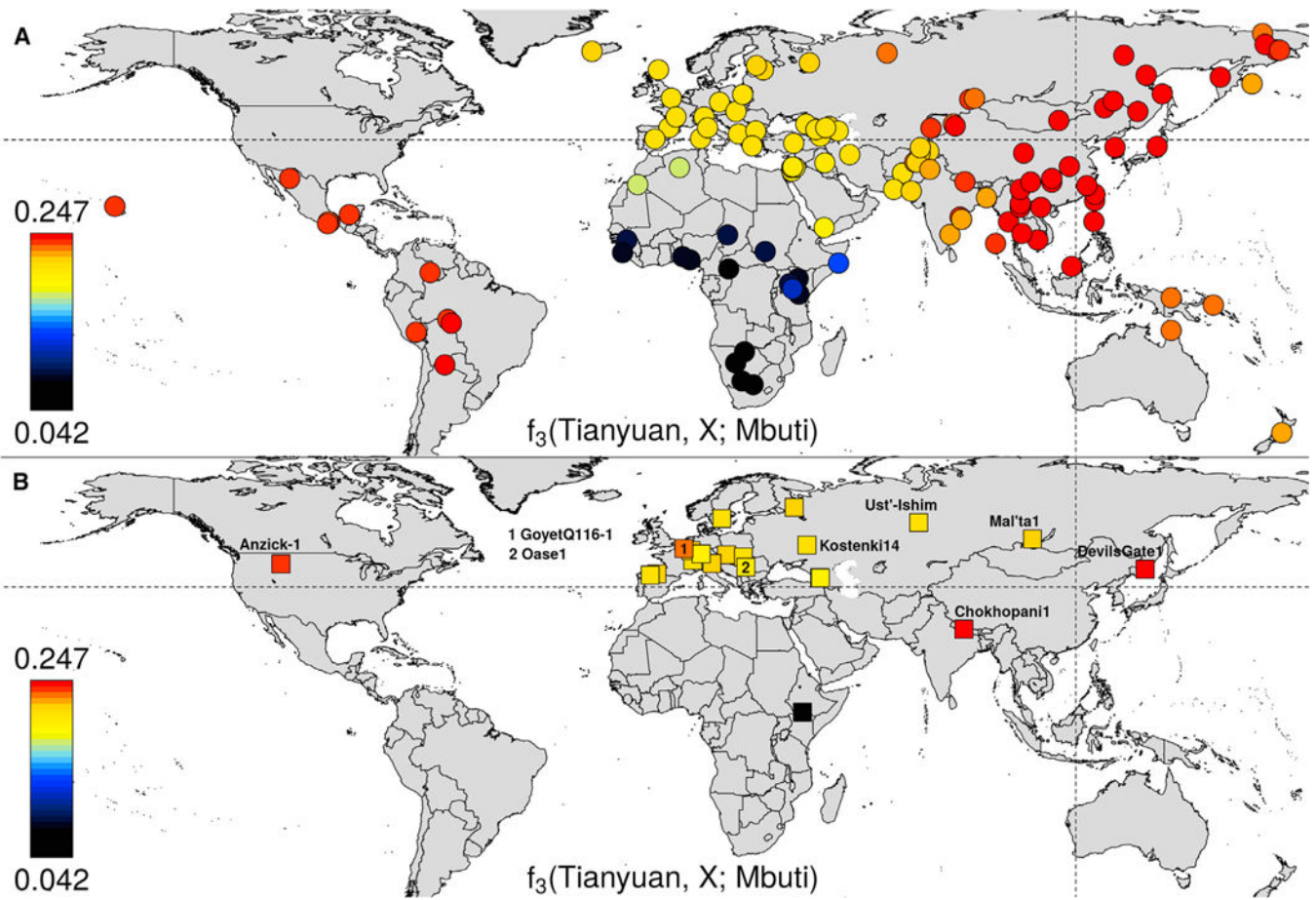
Figure 1. *f₃(Tianyuan, X; Mbuti)* f_3(*Tianyuan, X; Mbuti*) **for All Sites Where X Is a Present-Day Human Population or an Ancient Individual**

The f_3 statistic ranges from 0.04 to 0.25. A higher value (red) indicates higher shared genetic drift between the Tianyuan individual and the (A) present-day population or (B) ancient individual. The intersection of the dotted lines indicates where the Tianyuan Cave is located. See also Table S2A.
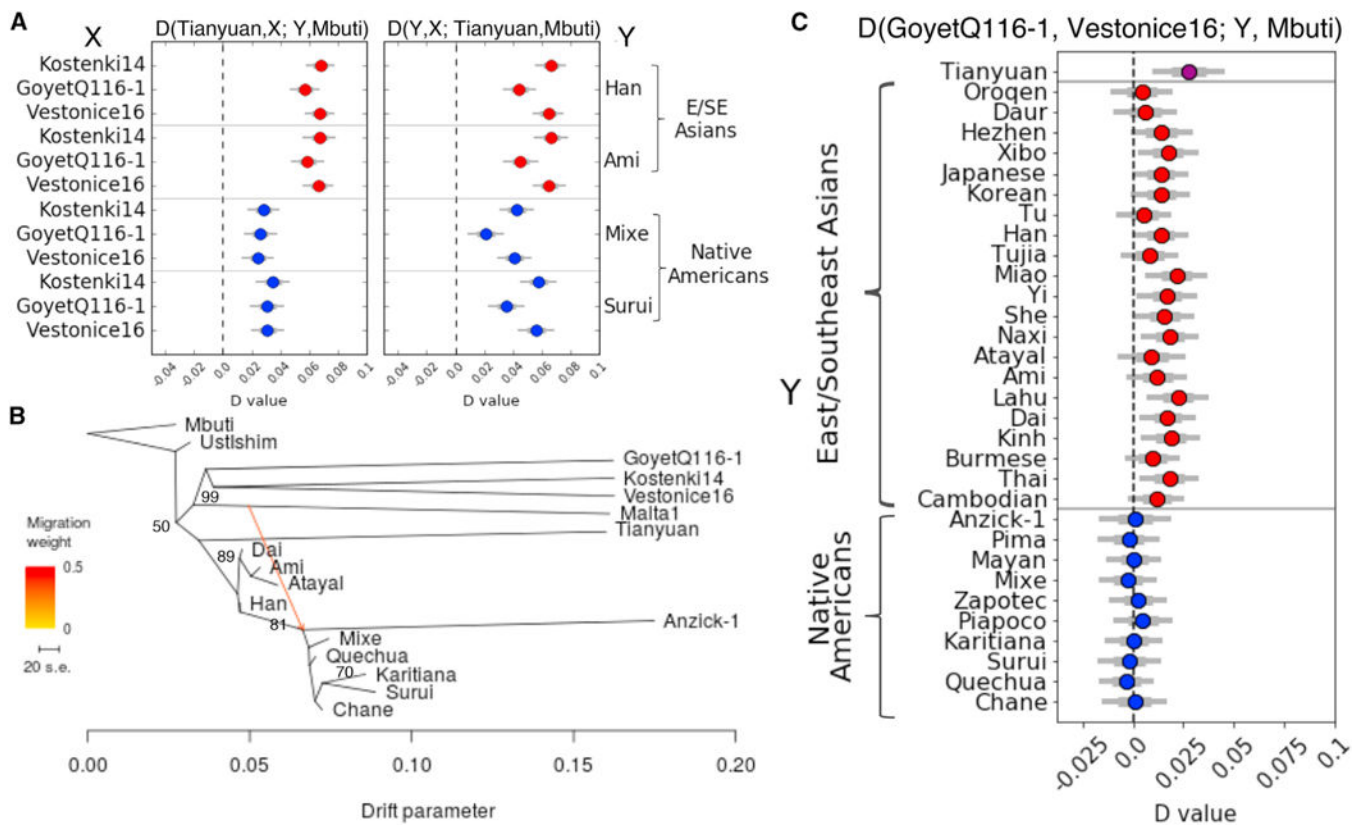
Figure 2. Comparisons Relating the Tianyuan Individual to Ancient and Present-Day Eurasians and Native Americans.

(A) *D(Tianyuan, X; Y, Mbuti)* and *D(Y, X; Tianyuan, Mbuti),* where Y is the East and Southeast Asian Han or Ami or the Native American Mixe or Surui and X is the ancient West Eurasian individual Kostenki14, GoyetQ116–1, or Vestonice16.

(B) Maximum-likelihood tree showing East and Southeast Asians, Native Americans, and ancient Eurasians, with bootstrap support of 100% unless indicated otherwise. The scale bar shows the average standard error (SE) of the entries in the covariance matrix.

(C) *D(GoyetQ116–1, Vestonice16; Y, Mbuti),* where Y is the Tianyuan individual, an East and Southeast Asian population, or a Native American population.

For (A) and (C), thick bars are within 1 SE of the estimate, thin bars are within 1.96 SE of the estimate (95% confidence interval), and the dashed vertical line indicates $D = 0$. See also Figures S1 and S2 and Tables S2 and S3.
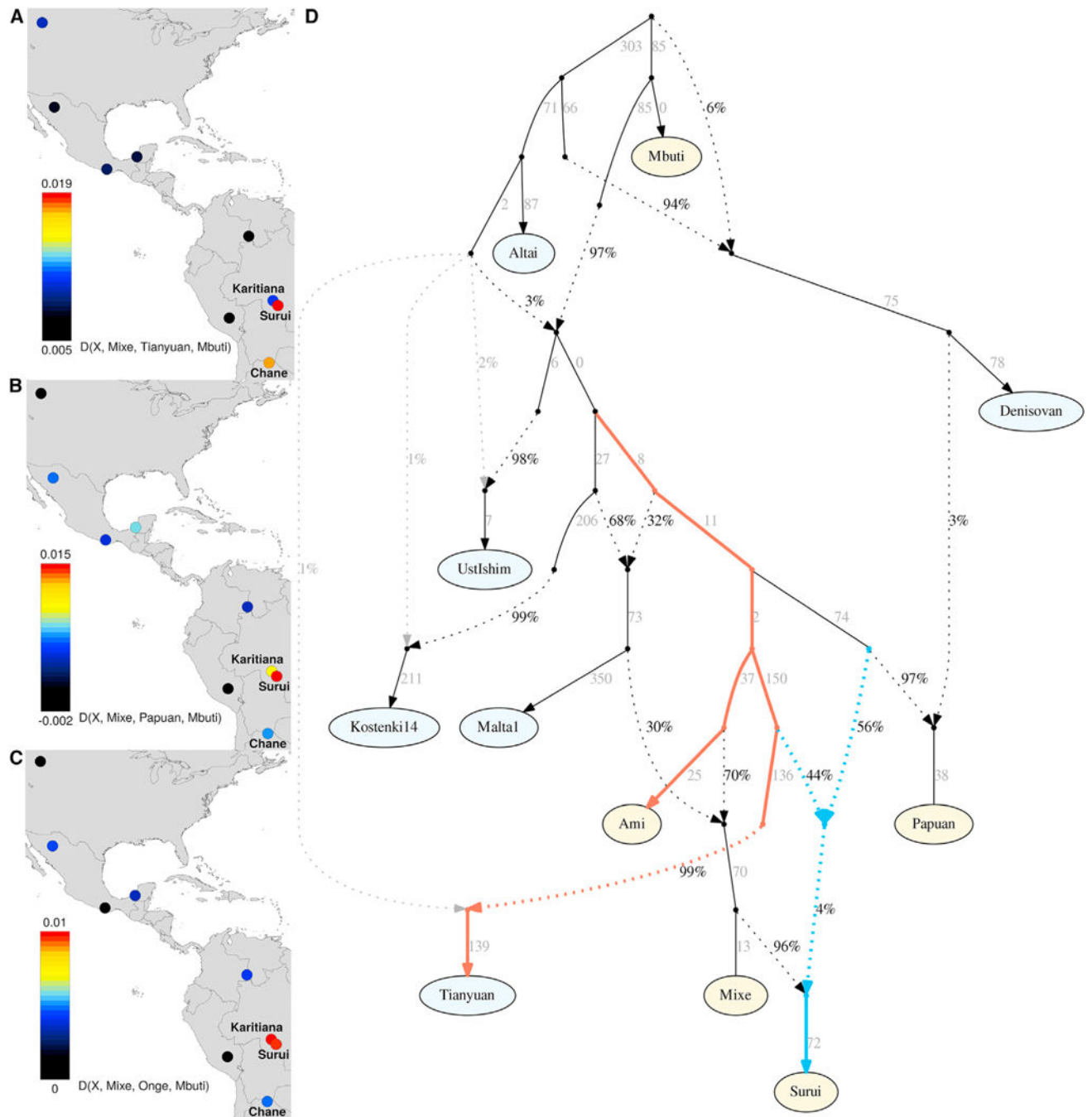
Figure 3. Heatmap of D Statistics and Admixture Graph Model Comparing Native Americans to the Tianyuan Individual.

(A–C) Heatmaps of (A) *D(X, Mixe; Tianyuan, Mbuti),* (B) *D(X, Mixe; Papuan, Mbuti),* and (C) *D(X, Mixe; Onge, Mbuti),* where X are non-Mixe Native American populations.

(D) Admixture graph model that fits allele frequency patterns (all empirical f statistics are within 3 SEs of expectation). Branch lengths are shown in units of $F_{st} \times 1,000$. Admixture from a population related to the Altai Neanderthal into ancient individuals (shaded gray) was collapsed into a single node, as were the original nodes at the top of the graph (can be observed in Figure S4N). The orange lines indicate the edges leading to the Tianyuan

individual and the East and Southeast Asian Ami after splitting from the edge leading to Kostenki14. The blue lines indicate the edges showing ancestral components related totheTianyuan individual and the Papuan in the Native American Surui.See also Figure S4 and Table S2.

Table 1.

Maximum-Likelihood Test Determining whether the Tianyuan Individual Is Directly Ancestral to any Present-Day Population

Population	c	k_1
Han	0.1800	0.2809
Dai	0.1808	0.2812
Miao	0.1818	0.2814
Japanese	0.1819	0.2813
Ami	0.1822	0.2813
Burmese	0.1834	0.2817
Hezhen	0.1838	0.2817
Oroqen	0.1841	0.2821
Lahu	0.1864	0.2828
Igorot	0.1866	0.2830
Thai	0.1883	0.2833
Cambodian	0.1887	0.2834
Kusunda	0.1972	0.2859
Mayan	0.1973	0.2860
Papuan	0.1975	0.2855
Karitiana	0.1992	0.2866
Australian	0.2001	0.2867
Uygur	0.2143	0.2921
French	0.2420	0.3013

The Tianyuan individual is not from a population directly ancestral to any of the listed present-day populations (c 0, where c refers to the amount of private drift for the Tianyuan individual). See STAR Methods for a description of c and k_1. See also Figure S3H.

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Biological Samples		
Tianyuan sample	Chinese Academy of Sciences Institute of Vertebrate Paleontology and Anthropology	See [7] and Tables S1A and S1B for more details
Chemicals, Peptides, and Recombinant Proteins		
Tween-20	Sigma	cat.no. P5927
1 M Tris-HCl, pH 8.0	AppliChem	A4577,0500
0.5 M EDTA, pH 8.0	AppliChem	A4892,1000
5 M NaCl	Sigma	S5150
25 mM each dNTP mix	Fermentas	cat.no. R1121
USER enzyme (mixture of Uracil-DNA-glycosylase, UDG, and Endonuclease, EndoVIII)	NEB	cat.no. M5505L
Bst DNA Polymerase, Large Fragment	NEB	cat.no. M0275S
ATP, 10 mM stock solution	NEB	cat. no. 9804
T4 Polynucleotide Kinase (10 U/μL)	NEB	cat. no. M0236L or S
T4 DNA Polymerase (3 U/μL)	NEB	cat. no. M0203L
Quick Ligation Kit	NEB	cat. no. M2200L
2xHI-RPM hybridization buffer	Agilent	5118–5380
20% SDS	Serva	39575.01
SSC buffer	Ambion	AM9770
AmpliTaq Gold 10x PCR buffer without Mg	Life Technologies	4379874
1M NaOH	Sigma	71463–1L
3M Sodium acetate pH 5.2	Sigma	S7899
Dynabeads MyOne C1	Life Technologies	65002
SeraMag Speedbeads	GE	6515210505250
Cot-1 DNA	Invitrogen	15279011
Deposited Data		

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Raw and analyzed data	This paper	ENA: PRJEB20217
Simons Genome Diversity Panel (SGDP)	[33]	RRID: SCR_006735; https://www.simonsfoundation.org/life-sciences/simons-genome-diversity-project-dataset/; see Table S1E
1000 Genomes Phase 3 dataset	[34]	RRID: SCR_008801; ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase3/data
Ancient individuals dataset	See Table S1D for original sources of compiled sequence data	N/A
Chimpanzee reference genome, PanTro2	[35]	http://genome.wustl.edu/pub/organism/Primates/Pan_troglodytes/
Oligonucleotides		
Probes for Panel 1, in [9], Supplemental Data 2a-d	[9]	See [6, 9] 'Methods' for more details
Probes for Panel 2, in [6], Supplemental Data 1	[6]	See [6] 'Methods' for more details
Probes for Panel 3, in [6], Supplemental Data 2	[6]	See [6] 'Methods' for more details
Probes for Panel 4, in [6], Supplemental Data 3	[6]	See [6] 'Methods' for more details
A*C*A*C*TCTTTCCCTACACGACGCTCTTCCG*A*T*C*T	N/A	NI1_P5_reg_short (=IS1)
G*T*G*A*CTGGAGTTCAGACGTGTGCTCTTCCG*A*T*C*T	N/A	NI2_P7_reg_short (=IS2)
A*G*A*T*CGGAA*G*A*G*C	N/A	NI3_P5P7_reg_comp (=IS 3)
A*C*A*C*TCTTTCCCTACACGACGCTCTTCCGATCT*A*C*T*C	N/A	NI4_P5_CR1_short
G*T*G*A*CTGGAGTTCAGACGTGTGCTCTTCCGATCT*A*C*T*C	N/A	NI5_P7_CR1_short
G*A*G*T*AGATC*G*G*A*A	N/A	NI6_P5P7_CR1_comp
A*C*A*C*TCTTTCCCTACACGACGCTCTTCCGATCT*G*T*C*T	N/A	NI7_P5_CR2_short
G*T*G*A*CTGGAGTTCAGACGTGTGCTCTTCCGATCT*G*T*C*T	N/A	NI8_P7_CR2_short
A*G*A*C*AGATC*G*G*A*A	N/A	NI9_P5P7_CR2_comp
GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT-Phosphate	N/A	BO4.P7.part1.R
CAAGCAGAAGACGGCATACGAGAT-Phosphate	N/A	BO6.P7.part2.R
GTGTAGATCTCGGTGGTCGCCGTATCATT-Phosphate	N/A	BO8.P5.part1.R
AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT-Phosphate	N/A	BO10.P5.part2.R
GGAAGAGCGTCGTGTAGGGAAAAGAGTGT-Phosphate	N/A	BO11.P5.part2.R
Software and Algorithms		
bwa	[36]	RRID: SCR_010910; http://bio-bwa.sourceforge.net/
ANGSD	[10]	http://popgen.dk/angsd/index.php/ANGSD

REAGENT or RESOURCE	SOURCE	IDENTIFIER
ADMIXTOOLS (*qp3Pop, qpDstat, qpF4ratio, qpWave, qpAdm, qpGraph*)	[9,11]	https://github.com/DReichLab/AdmixTools
Treemix	[37]	https://bitbucket.org/nygcresearch/treemix/wiki/Home
Test of Direct Ancestry Algorithm	This paper	N/A
ms	[38]	http://home.uchicago.edu/~rhudson1/source/mksamples.html