DNA analysis of an early modern human from Tianyuan Cave, China

Qiaomei Fu[a,b,1], Matthias Meyer[b], Xing Gao[a], Udo Stenzel[b], Hernán A. Burbano[b,c], Janet Kelso[b], and Svante Pääbo[a,b,1]

[a]Chinese Academy of Sciences–Max Planck Society Joint Laboratory for Human Evolution, Institute of Vertebrate Paleontology and Paleoanthropology, Chinese Academy of Sciences, 100044 Beijing, China; [b]Department of Evolutionary Genetics, Max Planck Institute for Evolutionary Anthropology, D-04103 Leipzig, Germany; and [c]Department of Molecular Biology, Max Planck Institute for Developmental Biology, D-72076 Tübingen, Germany

Contributed by Svante Pääbo, December 11, 2012 (sent for review September 21, 2012)

Hominins with morphology similar to present-day humans appear in the fossil record across Eurasia between 40,000 and 50,000 y ago. The genetic relationships between these early modern humans and present-day human populations have not been established. We have extracted DNA from a 40,000-y-old anatomically modern human from Tianyuan Cave outside Beijing, China. Using a highly scalable hybridization enrichment strategy, we determined the DNA sequences of the mitochondrial genome, the entire nonrepetitive portion of chromosome 21 (~30 Mbp), and over 3,000 polymorphic sites across the nuclear genome of this individual. The nuclear DNA sequences determined from this early modern human reveal that the Tianyuan individual derived from a population that was ancestral to many present-day Asians and Native Americans but postdated the divergence of Asians from Europeans. They also show that this individual carried proportions of DNA variants derived from archaic humans similar to present-day people in mainland Asia.

ancient DNA | human evolution | nuclear capture strategy | paleogenetics

The term "early modern humans" generally refers to humans who fall within the morphological variation of present-day humans and date to the Middle or Early Upper Paleolithic. The earliest modern humans appear in the Eurasian fossil record about 45,000 y ago, whereas the last remains that tend to be classified as early modern humans are about 25,000 y old (1). Early modern humans may exhibit some archaic features shared with other earlier forms of humans such as Neandertals. Although early modern humans are thus only vaguely defined as a group, their genetic relationship to present-day humans is unclear. Similarly, their relationship to archaic humans is of interest, given that they may have interacted directly with them.

To begin to explore the genetic relationships of early modern humans with present-day humans, we have analyzed a partial human skeleton that was unearthed in 2003, along with abundant late Pleistocene faunal remains, in the Tianyuan Cave near the Zhoukoudian site in northern China, about 50 km southwest of Beijing. The skeleton was radiocarbon-dated to 34,430 ± 510 y before present (BP) (uncalibrated), which corresponds to ~40,000 calendar years BP (2). A morphological analysis of the skeleton (3) confirms initial assessments (4) that this individual is a modern human, but suggests that it carries some archaic traits that could indicate gene flow from earlier hominin forms. The Tianyuan skeleton is thus one of a small number of early modern humans more than 30,000 y old discovered across Eurasia (2) and an even smaller number known from East Asia (5).

Results and Discussion

DNA Extraction. To evaluate DNA preservation and the degree of modern human contamination in the Tianyuan skeleton, we prepared two DNA extracts from the left femur (TY1301) and two from the right tibia (TY1305) of the human skeleton excavated in Tianyuan Cave (4) using less than 100 mg of bone material per extraction (Table S1). Sequencing of random DNA fragments from DNA libraries constructed from these four extracts revealed that about 0.01–0.03% of the DNA in the libraries was of human origin (*SI Text*, section 1 and Table S1). This low percentage of endogenous DNA precludes sequencing of the entire genome of this individual. Thus, we used DNA capture approaches to retrieve the mitochondrial (mt)DNA (6) and nuclear DNA sequences from the Tianyuan individual.

mtDNA Capture and Sequencing. We used a protocol for targeted DNA sequence retrieval that is particularly suited for mtDNA (6) to isolate mtDNA fragments. In total, we sequenced 4,423,607 unique DNA fragments from both ends on the Illumina GAII platform from the four libraries. Of these, 0.1–7.7% (Table S1) could be aligned to a human mtDNA reference sequence (7). All four libraries yielded consensus mtDNA sequences that agreed with each other. This is in agreement with the observation of paleontologists that the two bones come from a single individual (2). For estimation of mtDNA contamination and phylogenetic analyses, we used DNA fragments isolated from the femur library that had the highest mtDNA content (Table S1) and yielded an average coverage of 35.6-fold of the mtDNA genome.

DNA Sequence Authenticity. When studying modern humans, it is particularly difficult to exclude that DNA from present-day humans might have contaminated the samples or experiments. It has previously been argued (8) that the combination of two observations makes it likely that a DNA library contains a majority of endogenous ancient human DNA: First, that the patterns of DNA degradation, in particular nucleotide misincorporations resulting from deamination of cytosine residues at ends of DNA fragments, indicate that the mtDNA is ancient; and second, that deep sequencing of the mitochondrial genome indicates that the human mtDNA in the library comes from a single individual. We identified 78 distinct DNA fragments that cover three positions where the Tianyuan mtDNA consensus differs from at least 308 of 311 mtDNA genomes from around the world (9). All these fragments carry the consensus base at these positions, indicating that the vast majority [95% confidence interval (CI): 95.3–100%] of the endogenous mtDNA fragments come from a single source (*SI Text*, section 2). Alignments of the individual mtDNA fragments against the consensus sequence constructed from all fragments reveal C→T and G→A substitution frequencies of between 25 and 30% close to the ends of the DNA fragments

Author contributions: Q.F., H.A.B., J.K., and S.P. designed research; Q.F. performed research; M.M., X.G., U.S., and H.A.B. contributed new reagents/analytic tools; Q.F., M.M., U.S., J.K., and S.P. analyzed data; and Q.F., M.M., X.G., J.K., and S.P. wrote the paper.

The authors declare no conflict of interest.

Freely available online through the PNAS open access option.

Data deposition: The sequences reported in this paper have been deposited in the GenBank database (accession no. KC417443) and Sequence Read Archive (accession no. ERP002037).

[1]To whom correspondence may be addressed. E-mail: qiaomei_fu@eva.mpg.de or paabo@eva.mpg.de.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1221359110/-/DCSupplemental.

GENETICS

(Fig. S1*C*), an extent of substitution indicative of cytosine deamination not seen in present-day human contamination (8, 10). We conclude that the human DNA extracted from the Tianyuan skeleton comes from a single individual and is likely to be endogenous to the skeleton.

mtDNA Analyses. We estimated a phylogenetic tree for 311 modern human mtDNAs, the Tianyuan mtDNA, and a complete Neandertal mtDNA (Fig. S2). The Tianyuan mtDNA falls within the variation of present-day human mtDNA. It carries all substitutions that have been used to define a group of related mtDNA sequences —"haplogroup R"—and in addition a deletion of a 9-bp motif (5′-CCCCCTCTA-3′, revised Cambridge reference sequence positions 8,281–8,289) as well as a substitution at position 16,189 (Fig. S3*A*), which together have been used to define a group of related mtDNA sequences, "haplogroup B" (11–13), within haplogroup R. In addition, it carries four substitutions (5,348, 5,836, 11,257, 16,293) that are not defining subgroups of haplogroup B (14–17). Thus, it is related to the mtDNA that was ancestral to present-day haplogroup B (Fig. 1), which has been estimated to be around 50,000 y old (18) (50. 7 ka BP; 95% CI: 38.1–68.3 ka BP). We note that the age of the Tianyuan individual is compatible with this date.

Today, mtDNA of haplogroup B occurs in Native Americans, populations of the Russian Far East, Central Asia, Korea, Taiwan, Melanesia, and Polynesia. It is thus widespread in Asia and America. The fact that an individual who lived in the Beijing area 40,000 y ago carried a mitochondrial genome that is potentially ancestral to mtDNAs in all these areas suggests that there is at least some population continuity from the earliest modern humans in East Asia to present-day populations in these areas. However, although genetic reconstructions of human population histories in Eurasia have relied heavily on mtDNA variation (19–21), the extent to which this accurately reflects older population histories in the region is unknown. We therefore proceeded to analyze nuclear DNA sequences of this individual who represents a member of an early population in Asia.

Chromosome 21 Capture and Sequencing. The ability to sequence nuclear DNA sequences from the Tianyuan individual is limited by the fact that no more than 0.03% of the DNA extracted is endogenous to the bone. Previous studies (22–24) have shown that hybridization enrichment can be used to obtain nuclear DNA fragments from ancient samples. However, the commercially available hybridization systems used in these studies provide only a limited number of capture probes in each hybridization reaction. Because the enrichment of highly fragmented DNA requires large overlaps between probes, this limits the total size of genomic regions that can be targeted to a few megabases at best. To overcome this limitation, we modified a strategy previously described by Gnirke et al. (25) that uses oligonucleotides synthesized on arrays to construct probe libraries that are then amplified and converted into biotinylated RNA capture probes through in vitro transcription. Our approach differs in that first we use oligonucleotides from arrays with higher probe density (1 million probes per array); second, we exclude most of the linker sequence from oligonucleotide synthesis; third, we combine probe libraries from several arrays into a single "superprobe library"; and fourth, we use biotinylated DNA instead of RNA probes. This way, we produced a library consisting of 8.7 million different probes tiled at 3-bp intervals across the 29.8 Mbp of all nonrepetitive sequence in chromosome 21 (Fig. S4).

We produced additional DNA extracts and libraries from femur TY1301, this time using uracil DNA glycosylase and endonuclease VIII during library preparation to avoid nucleotide misincorporations induced by deaminated cytosine residues (26). From these libraries as well as the two initial libraries from the femur, we performed two successive enrichments for chromosome 21 fragments (Fig. S5). A total of 789,925 unique DNA fragments was identified that represents 1.75-fold coverage of the targeted regions. Separately, we also enriched for mtDNA by the same approach. This resulted in estimates of the average human mtDNA contamination in these libraries between 0.1% and 3.1% (*SI Text*, section 3.4). A comparison with shotgun

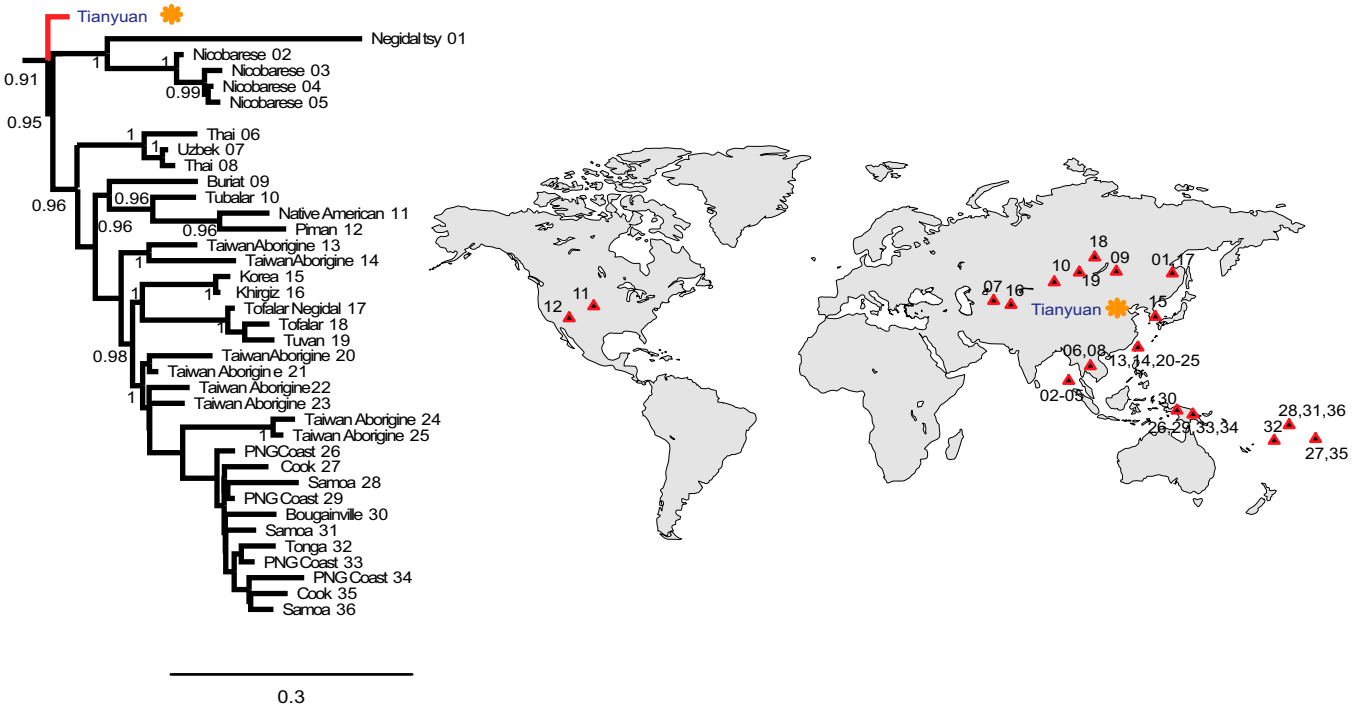

Fig. 1. Tree of the Tianyuan and 36 present-day mtDNAs belonging to haplogroup B. The bar represents 0.3 substitutions per nucleotide site. Numbers indicate individuals in the tree and the map.

sequencing results for two of the libraries used for capture showed that more than 70% of the total number of target DNA molecules present in the libraries was isolated by the capture procedure (*SI Text*, section 3.3). Thus, this procedure efficiently captures short molecules in very complex mixtures of DNA.

Chromosome 21 Analyses. To investigate the relationship of the Tianyuan individual to present-day populations, we compared it to chromosome 21 sequences from 11 present-day humans from different parts of the world (a San, a Mbuti, a Yoruba, a Mandenka, and a Dinka from Africa, a French and a Sardinian from Europe, a Papuan, a Dai, and a Han from Asia, and a Karitiana from South America) and a Denisovan individual, each sequenced to 24- to 33-fold genomic coverage (27). Denisovans are an extinct group of Asian hominins related to Neandertals (28). In the combined dataset, 86,525 positions variable in at least one individual are of high quality in all 13 individuals. Table 1 shows that the Tianyuan individual differs by 21,944–23,756 substitutions from the Eurasian individuals, by 30,297–35,938 substitutions from the African individuals, and by 43,893 substitutions from the Denisovan individual. Thus, the Tianyuan individual was clearly more similar to present-day humans than to Denisovans, and more similar to present-day Eurasians than to present-day Africans.

To more accurately gauge how the population from which the Tianyuan individual is derived was related to Eurasian populations, while taking gene flow between populations into account, we used a recent approach (29) that estimates a maximum-likelihood tree of populations and then identifies relationships between populations that are a poor fit to the tree model and that may be due to gene flow. As suggested by the nucleotide differences (Table 1), the maximum-likelihood tree (Fig. S6*A*) shows that the branch leading to the Tianyuan individual is long, due to its lower sequence quality. However, among Eurasian populations, Tianyuan clearly falls with Asian rather than European populations (bootstrap support 100%). The strongest signal not compatible with a bifurcating tree (Fig. S6*B*) is an inferred gene-flow event that suggests that 6.7% of chromosome 21 in the Papuan individual is derived from Denisovans, in agreement with previous findings (28, 30). When this is taken into account, the Tianyuan individual appears ancestral to all Asian individuals studied (Fig. 2). We note, however, that the relationship of the Tianyuan and Papuan individuals is not resolved (bootstrap support 31%). Further work is necessary to clarify whether this reflects the age of the Tianyuan individual relative to the divergence between modern human populations.

Archaic Admixture. It has been shown that a population related to the Denisovan individual contributed genetic material to the ancestors of present-day Melanesians (28, 30), and this has also been suggested to be the case for some mainland Asian populations (31) (but see also ref. 27). It is therefore of interest to analyze whether the Tianyuan individual shows evidence of any Denisovan genetic contribution. For chromosome 21, we find that any putative admixture from Denisovans must be smaller than that in present-day Papuans and not larger than in other present-day mainland Asians analyzed (*SI Text*, section 3.7 and Fig. S6*B*). However, because archaic admixture may show systematic differences among chromosomes (27), we decided to analyze additional parts of the Tianyuan genome for traces of archaic admixture. To do this, we identified 1,666 single-nucleotide polymorphisms (SNPs) where both the Neandertal (32) and Denisovan (27) genomes differ from the genomes of individuals from seven African populations in the Foundation Jean Dausset-Centre d'Etude du Polymorphisme Humain (CEPH)-Human Genome Diversity Panel (HGDP-CEPH) (33), and 1,800 SNPs where the Denisovan genome differs from the Neandertal as well as the seven Africans (*SI Text*, section 3.7). We synthesized capture probes for these 3,466 sites and generated additional DNA libraries from the Tianyuan individual. After enrichment and sequencing, we identified 834 and 843 sites, respectively, for which data were available for individuals from the HGDP-CEPH and the Tianyuan individual. Fig. 3 shows that all non-African populations share more alleles with the two archaic individuals. Over and above the alleles shared with the Neandertal, Melanesians share additional alleles with the Denisovan, as previously described (27, 28, 30), whereas the Tianyuan individual falls within the range of present-day Eurasian mainland populations. This indicates that the Tianyuan individual is most similar to the latter populations in carrying a genomic component related to the Neandertal genome, but no Denisovan component is discernable with these analyses (see also Fig. S7).

Conclusion

The DNA hybridization capture strategy described here allows sequencing of large sections (>>1 Mbp) of the nuclear genome from mammalian samples even in the presence of a large excess of microbial DNA, a situation typical of almost all ancient samples outside permafrost regions. This opens the possibility of generating DNA sequences from previously inaccessible ancient samples. We use this capture strategy to analyze an early modern human, the Tianyuan individual, who contains less than 0.03% endogenous DNA.

The results show that early modern humans present in the Beijing area 40,000 y ago were related to the ancestors of many present-day Asians as well as Native Americans. However, they had already diverged from the ancestors of present-day Europeans.

That Europeans and East Asians had diverged by 40,000 y ago is consistent with dates for the first archaeological appearance of modern humans in Europe and also with the upper end of an estimate [23 ka BP (95% CI: 17–43 ka BP)] for the divergence of East

Table 1. Pairwise nucleotide differences among chromosome 21 sequences analyzed

	San	Mbuti	Yoruba	Mandenka	Dinka	French	Sardinian	Papuan	Karitiana	Han	Dai	Tianyuan
Mbuti	33,860											
Yoruba	33,253	30,683										
Mandenka	33,783	31,489	26,684									
Dinka	33,207	31,099	27,624	27,906								
French	34,076	31,720	27,559	28,063	27,135							
Sardinian	34,138	31,936	27,901	27,331	27,245	18,326						
Papuan	34,672	32,754	29,023	28,871	28,931	22,630	21,968					
Karitiana	34,622	32,644	28,917	29,077	28,187	20,336	20,470	22,210				
Han	33,984	32,054	28,633	27,981	27,923	20,484	21,278	21,606	18,768			
Dai	34,173	32,209	28,092	28,332	27,704	20,211	21,683	21,297	20,131	18,525		
Tianyuan	35,938	33,390	31,059	30,333	30,297	23,168	22,906	23,756	21,944	22,802	23,339	
Denisovan	47,535	47,169	45,876	45,850	46,208	45,633	45,909	43,935	45,765	45,507	45,160	43,893

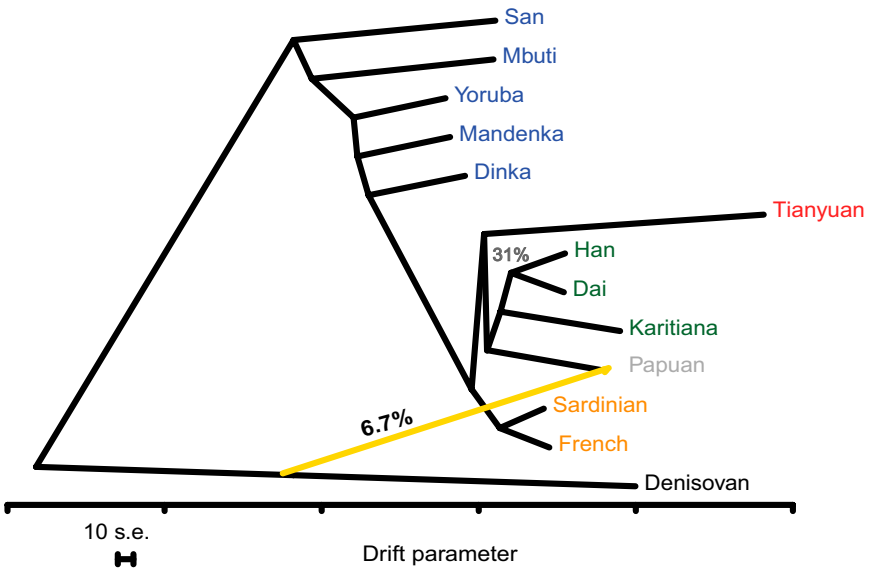

Fig. 2. Maximum-likelihood tree relating the chromosome 21 sequences of the Tianyuan individual, 11 present-day humans, and the Denisovan genome. The most strongly supported gene-flow event is shown in yellow. Bootstrap support for all internal edges is 100% except for the edge putting Tianyuan outside the four Asians, which is 31%. The scale bar shows 10 times the average standard error of the entries in the covariance matrix.

Asian and European populations from nuclear DNA variation in present-day populations (34). The results also show that the Tianyuan individual did not carry any larger proportion of Neandertal or Denisovan DNA sequences in its genome than present-day people in the region. More analyses of additional early modern humans across Eurasia will further refine our understanding of when and how modern humans spread across Eurasia.

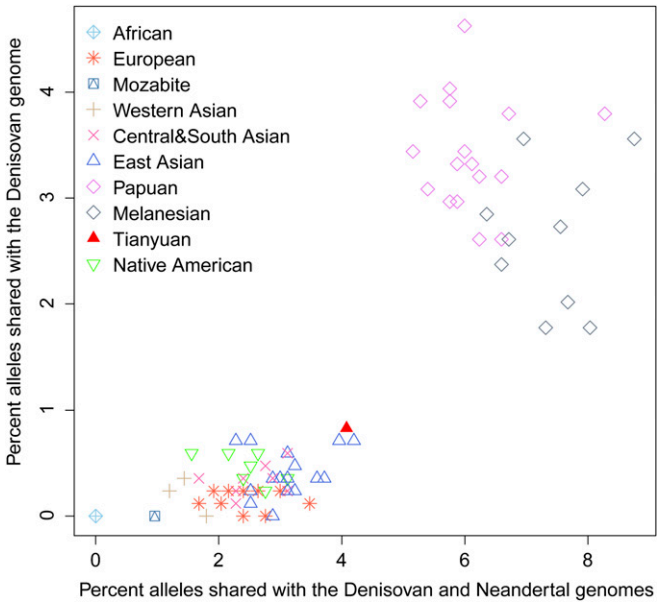

Fig. 3. Proportion of alleles shared with the Neandertal and Denisovan genomes for the Tianyuan and present-day individuals. The x axis shows the percent of alleles that match the Neandertal and Denisovan genomes at sites where both of these differ from seven Africans, and the y axis indicates the percent of alleles that match the Denisovan genome where this differs from the Neandertal as well as from the seven Africans.

Materials and Methods

DNA Library Preparation. The libraries used for mitochondrial capture and shotgun sequencing were produced from 20 μL of each extract as described (35) without enzymatic removal of deaminated cytosines (uracils) (Table S1). Both library adapters (P5 and P7) carried a 4-bp "clean-room key" sequence (5′-GTCT-3′) that is used exclusively for ancient DNA (36). In addition, sample-specific barcodes were introduced into both library adaptors during amplification (35), which was performed using AmpliTaq Gold DNA polymerase (Applied Biosystems).

For the chromosome 21 and SNP capture, additional extracts prepared from TY1301 were converted into sequencing libraries carrying the clean-room keys as described above. Miscoding DNA damage was removed during library preparation by treatment with uracil-DNA-glycosylase and endonuclease (EndoVIII) as described (26) (*SI Text*, section 3.1). Libraries were amplified with Herculase II Fusion DNA polymerase (Agilent) as described (37).

mtDNA Capture. mtDNA was enriched from the libraries separately using sheared human mtDNA from a European individual as bait as described (6).The enriched libraries were amplified using the primers IS5 and IS6 (38). After purification with Qiagen's MinElute PCR Purification Kit, library concentrations were determined on an Agilent 2100 Bioanalyzer DNA 1000 chip, pooled in equimolar amounts, and sequenced on a quarter of one lane of an Illumina Genome Analyzer IIx. The GenBank accession number of complete mtDNA genome is KC417443.

Nuclear DNA Capture. Nonrepetitive regions of chromosome 21, as well as polymorphic positions across the genome selected to detect archaic human admixture, were captured using single-stranded biotinylated capture probes prepared from a commercial array (*SI Text*, section 3.2). Libraries were reamplified and captures were performed twice (*SI Text*, section3.1). The Sequence Read Archive accession number of chromosome 21 as well as polymorphic position sequences is ERP002037.

mtDNA Phylogenetic Reconstruction. A Bayesian tree (Fig. S2) was estimated using MrBayes (39) with 50,000,000 Markov chain Monte Carlo iterations (5,000,000 burn-ins) using the Tianyuan consensus sequence, 311 human mtDNAs, and a Neandertal mtDNA (Vindija 33.25) (40). The general time-reversible sequence evolution model was applied with a fraction of invariable sites determined by the best-fit model approach of MODELTEST in conjunction with PAUP* (41). The haplogroup for each mtDNA was determined using Phylotree (Phylotree.org-mtDNA, build 15).

Chromosome 21 Sequence Determination. The Unified Genotyper from the Genome Analysis Toolkit was used to produce genotype calls, and a variant call format file combining the information with 13 other individuals was produced as described (27). We required the difference in Phred-scaled likelihoods between the two most likely genotypes to be at least 50 (corresponding to an error rate

of no more than 10^{-5}). When this did not result in a genotype call, we considered the two most likely homozygous genotypes and called the most likely one if their difference in Phred score was at least 50. The sites that are variable in 13 individuals were converted to TreeMix (29) input (*SI Text*, section 3.5). To compute pairwise distances, we restricted the analysis to 86,525 sites where a haploid or diploid genotype was called in all individuals, and not more than one nonreference allele was called across all individuals. The distance between two individuals at a site was defined as the difference in the number of reference alleles (Table S2), and distances were summed over all sites (Table 1).

ACKNOWLEDGMENTS. We thank Wu Xinzhi and Tong Haowen for their continual support, which made our work possible; Emily M. Leproust, Götz Frommer, and Leonardo Brizuela from Agilent Technologies for kindly providing special oligonucleotide arrays and technical advice; Martin Kircher and Birgit Nickel for invaluable technical help; Johannes Krause, Michael Lachmann, Daniel Lawson, Nick Patterson, Joseph Pickrell, David Reich, and Mark Stoneking for comments on the manuscript; and The Max Planck Society and its Presidential Innovation Fund, the Chinese Academy of Sciences Strategic Priority Research Program (Grant XDA05130202), and the Basic Research Data Projects (Grant 2007FY110200) of the Ministry of Science and Technology of China for financial support.

1. Trinkaus E (2007) European early modern humans and the fate of the Neandertals. *Proc Natl Acad Sci USA* 104(18):7367–7372.
2. Shang H, Tong H, Zhang S, Chen F, Trinkaus E (2007) An early modern human from Tianyuan Cave, Zhoukoudian, China. *Proc Natl Acad Sci USA* 104(16):6573–6578.
3. Shang H, Trinkaus E (2010) *The Early Modern Human from Tianyuan Cave* (Texas A&M Univ Press, College Station, TX).
4. Tong HW, Shang H, Zhang SQ, Chen FY (2004) A preliminary report on the newly found Tianyuan Cave, a Late Pleistocene human fossil site near Zhoukoudian. *Chin Sci Bull* 49(8):853–857.
5. Demeter F, et al. (2012) Anatomically modern human in Southeast Asia (Laos) by 46 ka. *Proc Natl Acad Sci USA* 109(36):14375–14380.
6. Maricic T, Whitten M, Pääbo S (2010) Multiplexed DNA sequence capture of mitochondrial genomes using PCR products. *PLoS One* 5(11):e14004.
7. Andrews RM, et al. (1999) Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA. *Nat Genet* 23(2):147.
8. Krause J, et al. (2010) A complete mtDNA genome of an early modern human from Kostenki, Russia. *Curr Biol* 20(3):231–236.
9. Green RE, et al. (2008) A complete Neandertal mitochondrial genome sequence determined by high-throughput sequencing. *Cell* 134(3):416–426.
10. Sawyer S, Krause J, Guschanski K, Savolainen V, Pääbo S (2012) Temporal patterns of nucleotide misincorporations and DNA fragmentation in ancient DNA. *PLoS One* 7(3):e34131.
11. Melton T, et al. (1995) Polynesian genetic affinities with Southeast Asian populations as identified by mtDNA analysis. *Am J Hum Genet* 57(2):403–414.
12. Redd AJ, et al. (1995) Evolutionary history of the COII/tRNALys intergenic 9 base pair deletion in human mitochondrial DNAs from the Pacific. *Mol Biol Evol* 12(4):604–615.
13. Hagelberg E, Cox M, Schiefenhövel W, Frame I (2008) A genetic perspective on the origins and dispersal of the Austronesians. *Past Human Migrations in East Asia: Matching Archaeology, Linguistics and Genetics*, Routledge Studies in the Early History of Asia, eds Sanchez-Mazas A, Blench R, Ross MD, Peiros I, Lin M (Routledge, New York), pp 356–375.
14. Kong QP, et al. (2006) Updating the East Asian mtDNA phylogeny: A prerequisite for the identification of pathogenic mutations. *Hum Mol Genet* 15(13):2076–2086.
15. Kivisild T, et al. (2002) The emerging limbs and twigs of the East Asian mtDNA tree. *Mol Biol Evol* 19(10):1737–1751.
16. Yao YG, Kong QP, Bandelt HJ, Kivisild T, Zhang YP (2002) Phylogeographic differentiation of mitochondrial DNA in Han Chinese. *Am J Hum Genet* 70(3):635–651.
17. Kumar S, et al. (2011) Large scale mitochondrial sequencing in Mexican Americans suggests a reappraisal of Native American origins. *BMC Evol Biol* 11:293.
18. Soares P, et al. (2009) Correcting for purifying selection: An improved human mitochondrial molecular clock. *Am J Hum Genet* 84(6):740–759.
19. Stoneking M, Delfin F (2010) The human genetic history of East Asia: Weaving a complex tapestry. *Curr Biol* 20(4):R188–R193.
20. Soares P, et al. (2010) The archaeogenetics of Europe. *Curr Biol* 20(4):R174–R183.
21. O'Rourke DH, Raff JA (2010) The human genetic history of the Americas: The final frontier. *Curr Biol* 20(4):R202–R207.
22. Burbano HA, et al. (2010) Targeted investigation of the Neandertal genome by array-based sequence capture. *Science* 328(5979):723–725.
23. Avila-Arcos MC, et al. (2011) Application and comparison of large-scale solution-based DNA capture-enrichment methods on ancient DNA. *Sci Rep* 1:Article 74.
24. Burbano HA, et al. (2012) Analysis of human accelerated DNA regions using archaic hominin genomes. *PLoS One* 7(3):e32877.
25. Gnirke A, et al. (2009) Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. *Nat Biotechnol* 27(2):182–189.
26. Briggs AW, et al. (2010) Removal of deaminated cytosines and detection of in vivo methylation in ancient DNA. *Nucleic Acids Res* 38(6):e87.
27. Meyer M, et al. (2012) A high-coverage genome sequence from an archaic Denisovan individual. *Science* 338(6104):222–226.
28. Reich D, et al. (2010) Genetic history of an archaic hominin group from Denisova Cave in Siberia. *Nature* 468(7327):1053–1060.
29. Pickrell JK, Pritchard JK (2012) Inference of population splits and mixtures from genome-wide allele frequency data. *PLoS Genet* 8(11):e1002967.
30. Reich D, et al. (2011) Denisova admixture and the first modern human dispersals into Southeast Asia and Oceania. *Am J Hum Genet* 89(4):516–528.
31. Skoglund P, Jakobsson M (2011) Archaic human ancestry in East Asia. *Proc Natl Acad Sci USA* 108(45):18301–18306.
32. Green RE, et al. (2010) A draft sequence of the Neandertal genome. *Science* 328(5979):710–722.
33. Patterson N, et al. (2012) Ancient admixture in human history. *Genetics* 192(3):1065–1093.
34. Gutenkunst RN, Hernandez RD, Williamson SH, Bustamante CD (2009) Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS Genet* 5(10):e1000695.
35. Kircher M, Sawyer S, Meyer M (2012) Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform. *Nucleic Acids Res* 40(1):e3.
36. Briggs AW, et al. (2007) Patterns of damage in genomic DNA sequences from a Neandertal. *Proc Natl Acad Sci USA* 104(37):14616–14621.
37. Dabney J, Meyer M (2012) Length and GC-biases during sequencing library amplification: A comparison of various polymerase-buffer systems with ancient and modern DNA sequencing libraries. *Biotechniques* 52(2):87–94.
38. Margulies M, et al. (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature* 437(7057):376–380.
39. Ronquist F, Huelsenbeck JP (2003) MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* 19(12):1572–1574.
40. Briggs AW, et al. (2009) Targeted retrieval and analysis of five Neandertal mtDNA genomes. *Science* 325(5938):318–321.
41. Posada D, Crandall KA (1998) MODELTEST: Testing the model of DNA substitution. *Bioinformatics* 14(9):817–818.

Supporting Information

Fu et al. 10.1073/pnas.1221359110

SI Text

1. SI Shotgun Sequencing

To determine the proportion of endogenous DNA, we carried out low-depth shotgun sequencing from the four libraries prepared from the Tianyuan skeleton. For this purpose, the barcoded libraries prepared from TY1301 (libraries B3071 and B3073) and TY1305 (libraries B3072 and B3074) were combined into two separate pools, each of which was sequenced on a lane of the Illumina Genome Analyzer IIx (FC-104-400x version 4 sequencing chemistry and PE-203-4001 cluster generation kit version 4) using a paired-end run with 76 + 7 cycles (1). An indexed control PhiX 174 library was spiked-in to yield 2–3% control reads (index 5′-TTGCCGC-3′). Base calling was performed with the machine-learning algorithm IBIS (2). Forward and reverse sequence reads overlapping by at least 11 bp were merged into single sequences to reconstruct full-length molecule sequences (3). These were used for further analysis. Merged reads were aligned against the human reference genome [National Center for Biotechnology Information (NCBI) accession no. 37/hg19] using BWA (4) with default parameters, and the output was converted to SAM/BAM format (5). The proportion of endogenous DNA in the four libraries ranges between 0.01% and 0.03% (Table S1), which makes the generation of sequences by whole-genome shotgun sequencing economically unfeasible. We therefore decided to proceed with hybridization enrichment.

2. SI Assembly of Mitochondrial DNA

Sequencing and processing of the raw data were performed as described for shotgun sequencing above. Mapping to the human mtDNA reference genome was done using an iterative mapping assembler (6) with a position-specific scoring matrix that takes into account the nucleotide misincorporation patterns found in ancient DNA sequences. To remove PCR duplicates, we built a consensus from sequences with identical start and end coordinates by retaining the base with the highest sum of quality scores at each position in the alignment. The average length of the mtDNA molecules is 59 bp (Fig. S1*A*). Mitochondrial coverage as determined from unique sequences ranges between 1.7- and 35.6-fold (Table S1). The consensus sequences obtained from the tibia and the femur are identical. The femur (TY1301) shows the better preservation of the two samples (0.48-fold mtDNA coverage per mg bone; Table S1). One of the libraries prepared from the femur (TY1301), which produced the highest coverage (35.6-fold), was used for further mtDNA analysis. To test whether the mtDNA fragments originated from one individual, the proportion of sequences that matched the consensus base at each position was calculated (Fig. S1*B*). The average support for the consensus base is 98.8%. Only 36 positions were covered with less than six sequences (the lowest threefold), but all sequences support the consensus base at these positions. The consensus support was below 80% for 8 out of 16,566 positions. However, manual inspection allowed a clear consensus call to be made for these eight cases: Five of the positions were incorrectly aligned, and another three positions showed more than one sequence with a C→T or G→A mismatch close to its end, suggesting that these substitutions represent nucleotide misincorporations due to cytosine deamination.

3. SI DNA Capture

3.1. DNA Library Amplification. To generate large quantities of amplified library for hybridization capture, the libraries prepared from the femur TY1301 for shotgun sequencing (B3071 and B3073), and a further 39 libraries where deaminated cytosines (uracils) had been enzymatically removed, were reamplified in 24 100-μL reactions using the primer pair "genomic R1" (5′-ACACTCTTTCCCTACACGACGCTCTTCCGATCT-3′) and "multiplex R2" (5′-GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT-3′).

Amplification products from each library were pooled and purified using solid-phase reversible immobilization (SPRI) technology (7) as follows: An SPRI solution was prepared by combining 95 g PEG 8000 (Promega), 50 mL 5 M NaCl, 2.5 mL 1 M Tris·HCl (pH 8.0), 0.5 mL 0.5 M EDTA (pH 8.0), and 125 μL Tween 20 and filling up to 250 mL with water. Five milliliters of carboxylated Sera-Mag Speedbeads (Distrilab BV) were washed twice in TE buffer, resuspended in 1 mL TE (10 mM Tris-HCl, 1 mM EDTA, pH 8.0), and added to the SPRI solution to obtain a ready-to-use 38% (wt/vol) PEG-SPRI suspension, which was stored in the refrigerator until used. Pooled PCR products from each library were mixed with an equal volume of SPRI suspension and incubated for 30 min at room temperature. Beads were pulled to the tube wall using a magnet rack and washed twice with 70% ethanol. After drying for 30 min, DNA was eluted in 60 μL TE buffer. The concentrations of the amplified libraries were between 1,036 and 1,268 ng/μL as determined on a NanoDrop 1000 photospectrometer. The original two libraries (B3071 and B3073) and five of the new libraries were captured twice using probes for chromosome 21 and twice using probes for mtDNA. A set of 34 of the new libraries was captured twice using the probes for the admixture SNPs.

3.2. Target Enrichment. ***3.2.1. Chromosome 21 capture probe design.*** Using the human reference genome sequence (hg19), 11,701,943 probe sequences of 52 bases were extracted from chromosome 21 with 3-bp tiling. To eliminate repetitive sequences, probes containing 15-mer sequences that are overrepresented in the human genome were removed (8). This resulted in the removal of ∼25% of the tiled probes, and 37,159 contiguous regions covered by probes remained (Fig. S4). Approximately 35 Mbp of sequence are present in the hg19 chromosome 21 reference assembly. Of these, 29.8 Mb (85%) were targeted by at least one probe, and 22.7 Mb (65%) were covered by 17 or 18 probes. A universal flanking sequence (5′-CACTGCGG-3′) was attached to the 3′ end of each of the 8,722,911 probe sequences. The resulting 60-base probes were printed on nine custom-designed 1 million-feature arrays (Agilent).

An additional array with 5,506 probes targeting the complete human mitochondrial genome (hg19 as provided by the University of California Santa Cruz Browser, http://genome.ucsc.edu/cgi-bin/hgGateway) using 3-bp tiling was designed as above. We removed 38 repetitive probes using the 15-mer filtering. The complete set of 5,468 probes that passed the filtering was printed 178 times in a 1 million-feature array generating a total of 973,304 probes.

3.2.2. "Admixture SNP" capture probe design. To further investigate what proportion of archaic admixture might be present in Tianyuan, we identified sites where seven individuals from seven different African populations (Bantu Kenya, Bantu South Africa, Biaka, Mandenka, Mbuti, San, and Yoruba) from the CEPH-Human Genome Diversity Panel (HGDP-CEPH) differ from both the Denisovan and Neandertal genomes (set A) and sites where all these African individuals match the Neandertal genome but are all different from the Denisovan (set B) (9). We designed a capture array for the 1,666 sites in set A and the 1,800 sites in set B.

Using the human reference genome sequence (hg19), we designed 52-bp–long probes tiled at 3-bp intervals across 105 bp

centered on each SNP. For each probe, alternatives carrying each of the two allelic variants of the SNP were included. To eliminate repetitive sequences, 49% of the probes containing 15-mer sequences that are overrepresented in the human genome were removed (8). The sequence 5′-CACTGCGG-3′ was attached to the 3′ end of each of the 124,163 probes. The resulting 60-base probes were printed on nine custom-designed 1 million-feature arrays (Agilent).

3.2.3. Generation of biotinylated capture probes. The array probes were cleaved and converted into probe libraries as follows. After adding 500 μL elution solution (125 mM NaOH, 0.05% Tween 20), each array was assembled with a gasket slide in a hybridization chamber (8) and rotated for 6 h with 12 rpm at room temperature. The eluate was recovered using a syringe, neutralized by adding 19 μL 20% acetic acid, and purified using the QIAquick Nucleotide Removal Kit (Qiagen). The purified probes were eluted in 20 μL EB (10 mM Tris-Cl, pH 8.5). Using 2 μL of the eluate, successful probe recovery was confirmed by denaturing PAGE. The first probe library adapter was added through a primer extension reaction with Bst polymerase. The 50-μL reaction mixture contained 10 μL of purified probes and 3 μL of (24 U) Bst polymerase (large fragment; New England BioLabs), and in final concentrations 1× Thermopol buffer (New England BioLabs), 250 μM each dNTP, and 1 μM extension primer APL2 (5′-biotin-CGTGGATGAGGAGCCGCAGTG-3′). After incubation for 1 min at 50 °C, 5 min at 15 °C, 5 min at 20 °C, 5 min at 25 °C, 5 min at 30 °C, and 10 min at 37 °C in a thermal cycler, the reaction was purified using the MinElute PCR Purification Kit (Qiagen). Subsequently, a blunt-end repair reaction was performed to remove 3′ overhangs generated by Bst polymerase. The 40-μL reaction mixture contained the complete eluate (20 μL), 0.4 μL (2 U) T4 DNA polymerase (Fermentas), and 0.4 μL (4 U) Klenow fragment (Fermentas), and in final concentrations 1× Tango buffer (Fermentas) and 100 μM each dNTP. After incubation for 15 min at 25 °C, the reaction was purified using the MinElute PCR Purification Kit, eluting in 20 μL EB. The second probe library adapter was added by blunt-end ligation as follows. A double-stranded adapter was generated by combining 7 μL water, 1 μL T4 DNA ligase buffer (Fermentas), 1 μL 100 μM APL1 (5′-phosphate-ACACGCTGGTGCGATCC-CTAT-Pho-3′), and 1 μL 100 μM APL6 (5′-ATAGGGATCGC-ACCAGCGTGT-3′). The mixture was incubated for 10 s at 95 °C in a thermal cycler and slowly cooled to 14 °C at a rate of 0.1°/s. Then, 3 μL T4 DNA ligase buffer, 2 μL water, 4 μL 50% PEG 4000, and the eluate from the previous reaction were added. After mixing, 1 μL (5 U) T4 DNA ligase (Fermentas) was added and the reaction was incubated for 30 min at room temperature and purified using the MinElute PCR Purification Kit. To remove adapter dimers, 20 μL MyOne C1 streptavidin beads (Invitrogen) were washed twice with BWT+SDS buffer (1 M NaCl, 10 mM Tris·HCl, pH 8.0, 1 mM EDTA, 0.05% Tween 20, 0.5% SDS). The beads were resuspended in 180 μL BWT+SDS buffer, the complete eluate from the ligation reaction was added (20 μL), and the suspension was rotated at room temperature for 20 min. The beads were washed twice with 0.1× BWT buffer (0.1 M NaCl, 10 mM Tris·HCl, pH 8.0, 1 mM EDTA), resuspended in 25 μL TT buffer (1 mM Tris·HCl, pH 8.0, 0.01% Tween 20), and incubated for 10 min at 95 °C to release the biotinylated strands, representing the final probe library. Using 1 mL of the probe library and a standard dilution series of known concentration, the number of probe library molecules was determined by quantitative (q)PCR using the primer pair APL5–APL6. Based on this assay, we estimated that on average 16,000 copies of each probe were recovered.

Probe libraries were amplified for nine cycles in 100-μL PCR reactions (avoiding PCR plateau). Each reaction contained 10 μL probe library (~6,000 copies per probe) and 1 μL Herculase II Fusion DNA polymerase (Agilent), and in final concentrations 1× Herculase II reaction buffer, 250 μM each dNTP, and 400 nM the primers APL5 (5′-CGTGGATGAGGAGCCGCAGTG-3′) and APL6. An initial denaturation step of 2 min at 95 °C was followed by nine cycles of denaturation at 95 °C for 20 s, annealing at 60 °C for 30 s and elongation at 72 °C for 30 s, and a final extension step at 72 °C for 5 min. PCR products were purified using the MinElute PCR Purification Kit and eluted in 20 μL EB. Seven microliters of each product was then loaded on a 3% low-melting/1% high-melting agarose gel with SYBR Safe (Invitrogen). Narrow bands around 94 bp were excised from the gel to remove a faint smear of below–full-length probes. DNA was isolated from the gel slices using the MinElute Gel Extraction Kit, eluting in 30 μL EB. One microliter of each eluate was used for qPCR to verify the success of gel extraction and determine an optimal cycle number for the subsequent amplification (avoiding PCR plateau). Between 8 and 19 μL of gel-excised probes was used as template for amplification reactions in 100-μL volumes, containing 1 μL Herculase II Fusion DNA polymerase and in final concentrations 1× Herculase II reaction buffer, 250 μM each dNTP, and 400 nM the primers APL5 and APL4 (5′-GGATTCTAATACGACTCACTATAGGGATCGCACCAGC-GTGT-3′). An initial denaturation step of 2 min at 95 °C was followed by eight cycles of denaturation at 95 °C for 20 s, annealing at 60 °C for 30 s and elongation at 72 °C for 30 s, and a final extension step at 72 °C for 5 min. The PCR products were purified using the MinElute PCR Purification Kit, eluted in 40 μL EB, and quantified using a NanoDrop photospectrometer. Concentrations of the amplified probe libraries varied between 69 and 97 ng/μL. The nine probe libraries for chromosome 21 were pooled in equimolar ratio. At this stage, the probe library contains a T7 promotor sequence (introduced by APL4), in principle allowing for the generation of RNA capture probes following Gnirke et al. (10). However, because we did not observe substantial differences in the performance of DNA and RNA probe capture in earlier experiments, we chose to perform all captures of the Tianyuan libraries with DNA probes.

Single-stranded biotinylated DNA probes were generated in single-primed linear amplification reactions using a biotinylated primer. Because high amounts of template were required for these reactions, the chromosome 21 and mtDNA probe libraries were further amplified using the primer pair APL2 and APL6 under the conditions described above. For each probe set, 96 100-μL reactions were prepared, containing 200 μL template and 2 μL Herculase II Fusion DNA polymerase, and in final concentrations 1× Herculase II reaction buffer, 250 μM each dNTP, and 400 nM APL2. An initial denaturation step of 2 min at 95 °C was followed by 20 cycles of denaturation at 95 °C for 20 s, and annealing at 60 °C for 20 s and elongation at 72 °C for 20 s. All reactions were pooled in a Falcon tube and mixed with a double volume of 38% PEG-SPRI suspension (~20 mL). All other steps of the SPRI purification were performed as described above. Probes were eluted in 150 μL TE buffer and quantified using a NanoDrop photospectrometer (~250 ng/μL). DNA probes were stored at −20 °C until used.

3.2.4. Hybridization capture. For each hybridization reaction, a sample library pool (15 μL total volume) was created by combining 6 μL sample library (~2 μg), 5.25 μL water, 2.5 μL 1 mg/mL human Cot-1 DNA (Invitrogen), 0.25 μL 10 mg/mL salmon sperm DNA (Invitrogen), 0.5 μL 500 μM BO4 (5′-GTGACTGGAGTTCA-GACGTGTGCTCTTCCGATCT-phosphate-3′), and 0.5 μL 500 μM BO10 (5′-AGATCGGAAGAGCGTCGTGTAGGGAAAGAG-TGT-phosphate-3′). The sample library pool was incubated for 5 min at 95 °C and then 5 min at 65 °C, and held at room temperature afterward. A probe pool was created by diluting 300 ng single-stranded DNA probes with water to obtain a total volume of 4 μL. Hybridization buffer was prepared by combining 1.7 mL Hi-RPM buffer (aCGH Hybridization Kit; Agilent) and 300 μL 50× Denhardt's solution (Sigma-Aldrich). Hybridization re-

actions were assembled in 96-well plates by adding 20 μL hybridization buffer and the complete sample library pool to the probe pool. The reactions were mixed and incubated at 62 °C for 2 d. For each reaction, 30 μL MyOne T1 streptavidin beads (Invitrogen) were washed once with 150 μL wash buffer 1 (1× SSC, 0.01% SDS) for 15 min at room temperature, three times with 120 μL HWT buffer [1× AmpliTaq Gold buffer without $MgCl_2$ (Applied Biosystems), 0.02% Tween 20] for 10 min at 60 °C, and once with 150 μL wash buffer 3 (0.1 × SSC, 0.05% Tween-20) for 5 min at room temperature. Before each wash step, liquid was collected in the bottom of the wells by briefly spinning the plate at 2,000 × *g* in a centrifuge. The plate was then placed on a 96-well ring magnet plate and the supernatant was removed. Wash buffer was added and the plate was sealed with strip caps. Beads were fully resuspended by vortexing for 5–10 s. During room-temperature wash steps, the plate was taped to a rotator. For high-temperature wash steps, the plate was placed in a thermal cycler (with lid heating turned off) and inverted several times during incubation. After the last wash step, to eluate the capture library molecules, the beads were resuspended in 19 μL melt solution (125 mM NaOH, 0.05% Tween 20). After incubation for 15 min at room temperature, the supernatant was transferred to a fresh plate and mixed with 0.7 μL 20% acetic acid and 190 μL PN buffer (Qiagen). Each capture eluate was then purified in a separate MinElute spin column following Qiagen's instructions for using the Nucleotide Removal Kit. DNA was eluted in 30 μL TT buffer.

To verify the successful retrieval of library molecules, molecule numbers were estimated from 1 μL of capture eluate by qPCR (~1E8 total molecules for most libraries). The remaining 29 μL was amplified in 100-μL reactions, containing 1 μL Herculase II Fusion DNA polymerase, and in final concentrations 1× Herculase II reaction buffer, 250 μM each dNTP, and 400 nM the primers genomic R1 and multiplex R2. An initial denaturation step of 2 min at 95 °C was followed by 28 cycles of denaturation at 95 °C for 30 s, annealing at 60 °C for 30 s and elongation at 72 °C for 30 s, and a final extension step at 72 °C for 5 min. Amplified libraries were purified using a 1:1 ratio of 38% PEG-SPRI suspension as described above. The amplified capture eluates were eluted in 15 μL TE and their concentrations were determined using a NanoDrop photospectrometer (~170 ng/μL on average).

Three microliters (~500 ng) of capture eluate from the first round of hybridization was used as template for a second round of hybridization, which was performed under the same conditions except for a reduced incubation time (1 d). Capture eluates were again quantified by qPCR (values were one to two orders of magnitude higher), amplified (23 cycles), purified with SPRI beads, and quantified on a NanoDrop photospectrometer.

3.2.5. Library pooling and multiplex sequencing. The amplified capture eluates were diluted 25-fold using TE buffer (to obtain concentrations of ~5 ng/μL). To enable highly accurate multiplex sequencing (1), indexes were added to both library adapters using 5′-tailed PCR primers in 50-μL reactions containing 1 μL template and 0.5 μL Herculase II Fusion DNA polymerase, and in final concentrations 1× Herculase II reaction buffer, 250 μM each dNTP, and 400 nM each indexing primer. Amplification was performed with a small cycle number to avoid the formation of heteroduplices in PCR plateau. An initial denaturation step of 2 min at 95 °C was followed by six cycles of denaturation at 95 °C for 30 s, annealing at 60 °C for 30 s and elongation at 72 °C for 30 s, and a final extension step at 72 °C for 5 min. Fifty microliters of PB buffer (Qiagen) and 1 μL 3 M sodium acetate were added to each reaction. By pooling the sample/PB mixes in equimolar ratios, several library pools were generated. Libraries from chromosome 21 and mtDNA capture were kept in separate pools. After adding the twofold volume of PB buffer to each pool, the libraries were purified using the MinElute PCR Purification Kit and eluted in 25 μL TE. DNA concentration was determined using a DNA 1000 chip on a Bioanalyzer 2100 (Agilent).

Sequencing and raw sequence processing were performed as described in SI Text, section 1. We generated one lane of sequence data from the chromosome 21 captures. mtDNA captures were sequenced on one-quarter and the SNP capture was sequenced on one-fifth of an Illumina Genome Analyzer IIx lane. Alignments were generated by mapping the merged reads against the human reference genome (NCBI accession no. 37/hg19) using BWA (4) with default parameters.

3.3. Evaluating the Efficiency of the Target Enrichment Method. To evaluate whether the enriched libraries had been sequenced to exhaustion, we performed a subsampling analysis by randomly drawing subsets of aligned sequences and counting the number of unique sequences in each subset based on identical alignment start and end coordinates. Because capture had been performed in replicates for each library, we performed this analysis separately for the sequences from each replicate as well as the combined sequences from both replicates (Fig. S5). The subsampling curves provide two insights. First, sequencing depth was sufficient to almost completely exhaust the complexity of the enriched libraries, that is, deeper sequencing would not considerably increase the number of sequences of unique DNA fragments. Second, the trajectories of the subsampling plots are virtually identical, irrespective of whether sequences from capture replicates are analyzed separately or in combination. This indicates that the number of sequences obtained from unique DNA fragments is merely dependent on sequencing depth and that replicate captures from the same libraries are not required.

We next wanted to assess what proportion of the target molecules existing in the libraries we successfully captured and sequenced via hybridization enrichment. For this purpose, we made use of the whole-genome shotgun sequences that had been generated from some of the libraries (B3071 and B3073) that were also used for capture. We first searched for sequences that aligned to the capture target regions of chromosome 21 in the whole-genome shotgun data and identified 52 such sequences. Based on identical start and end alignment coordinates, we next checked how many of these sequences were also present in the chromosome 21 capture data. For B3071, we found that 18 out of 23 (78%) DNA fragments with a length of ≥35 bp and a mapping quality of ≥30 are also represented in the capture dataset. For B3073, we determined a similar number (74%; 14 out of 19 sequences). These results demonstrate that the capture strategy presented in this study efficiently captures most of the target molecules present in a library.

3.4. Chromosome 21 Sequence Coverage and mtDNA Contamination Estimate. In total, we obtained 9,373,365 sequences from five uracil-DNA-glycosylase (UDG)–treated libraries with a length of at least 35 bp, 4,406,261 of which (46.8%) aligned to chromosome 21 with a mapping quality of at least 30. For each library, sequences that map to the same outer reference coordinates were replaced by the sequence with the highest sum of base qualities (11). Through this approach, 789,925 sequences were found to be unique, and each unique molecule was sequenced on average 5.6 times. When combining sequences from all five libraries, 19.9 Mbp of 29.8 Mbp of the Tianyuan chromosome 21 target regions were captured, and average coverage is 1.75-fold.

To estimate human mitochondrial contamination in each of the libraries used for nuclear DNA capture, we focused on three positions where the Tianyuan mitochondrial consensus sequence differs from at least 99% of 311 present-day human mitochondrial genomes (12) (position 5,348: C→T; position 5,836: A→G; position 11,257: C→T) (Fig. 3*B*). The positions 5,348: C→T and 11,257: C→T may appear to be due to deamination; however, this seems unlikely, because (*i*) the libraries are UDG-treated, (*ii*) the

majority of reads agree on the nucleotide, and (*iii*) these positions are not near the ends of the majority of reads, as is typical for deaminations. We then counted the unique fragments that cover these positions to determine how many differ from the consensus sequence, that is, are putative contaminants. We estimate human mitochondrial contamination across all libraries to be on average 1.1% (highest, 3.1%; lowest, 0.1%).

3.5. Chromosome 21 Sequence Determination. We disregarded sites where all 13 individuals have the same genotype; 171,853 sites are variable in at least one of among the 13 individuals. To ensure that genotyping errors do not dominate the differences between individuals, we required the difference in Phred-scaled likelihoods between the two most likely genotypes to be at least 50 (corresponding to an error rate of no more than 10^{-5}). At sites where we could not call a diploid genotype in this way, we considered whether the likelihoods of the two most likely homozygous genotypes differ by at least 50, and if so called the most likely haploid genotype instead. A total of 87,243 sites pass this filter.

To also use the information from the sites called as haploid genotypes in subsequent analyses, haploid calls were counted as zero or one reference allele out of one observation, whereas the diploid calls were counted as zero, one, or two reference alleles out of two observations.

To compute pairwise distances, we ignored sites where two or more alternative alleles were seen among the individuals analyzed. A total of 86,525 sites passed this filter. The distance between two genotypes was defined as the difference in the number of reference alleles (Table S2). To calculate the distance between two individuals, the distances between genotypes were summed over all sites.

3.6. Phylogenetic Reconstruction of Chromosome 21. TreeMix estimates a tree (Fig. S6*A*) where there is 100% bootstrap support for Tianyuan clustering together with Karitiana, Han, and Dai (100 bootstrap replicates), irrespective of whether an admixture event between the Denisovan and Papuan populations is taken into account (Fig. 2 and Fig. S6*A*).

3.7. Archaic Admixture. We note that the distance between the Denisovan and the Tianyuan chromosome 21 sequences (43,893) is similar to the distance between the Denisovan and the Papuan sequences (43,935) and smaller than the distances between the Denisovan and the other Asian sequences (45,160–47,535) (Table 1). We therefore explored whether a population related to the Denisovan individual may have contributed genes to the ancestors of the Tianyuan individual, as is the case with present-day Melanesians (13, 14) and has been suggested for some mainland Asian populations (15), although this is more likely to represent Neandertal gene flow (16). Although the residuals (Fig. S6*B*) from the TreeMix analysis (which can be interpreted as signals of admixture) detect the previously described admixture signal between Denisovans and Papuans (Fig. 2), there is no indication of admixture between Denisovans and the Tianyuan individual that exceeds that seen between Denisovans and any other population analyzed.

To further investigate what proportion of archaic admixture might be present in Tianyuan, we identified sites where seven individuals from seven different HGDP-CEPH Africans (Bantu Kenya, Bantu South Africa, Biaka, Mandenka, Mbuti, San, and Yoruba) differ from both the Denisovan and Neandertal genomes (set A), and sites where all of these African individuals match the Neandertal genome but differ from the Denisovan genome (set B). The sharing of alleles seen in both archaic genomes (set A) tends to detect archaic human admixture both from Neandertals and Denisovans or groups related to them, whereas sharing of alleles seen in the Denisovan but not the Neandertal (set B) tends to show admixture with Denisovans or groups related to them.

We randomly selected one individual from each of the 44 populations in mainland Eurasia and the Americas represented in the HGDP-CEPH and counted the number of alleles each individual shares with the archaic individuals in the two sets of SNPs (Fig. 3). We similarly analyzed the 11 sequenced present-day humans and the Tianyuan. Here 1,499 sites in set A and 1,618 sites in set B were scored in all individuals analyzed (Fig. S7). We find no evidence for Denisovan gene flow into the population from which the Tianyuan individual is derived, above what might be present in all mainland Asian populations.

We note that there are at least two potential explanations for the relatively small distance between the Tianyuan individual and the Denisovan individual other than Denisovan gene flow. First, Denisovan and Tianyuan might share some artifacts present in ancient DNA sequences. However, whereas errors are present in the Tianyuan sequences due to its low coverage, the Denisovan genome sequence is of ~30-fold coverage and the DNA fragments are sequenced to such a high redundancy that its error rate is lower than the present-day sequences analyzed here (16). Second, the smaller distance between the Tianyuan and Denisovan individuals may be due to the fact that the 40,000-y-old Tianyuan individual as well as the perhaps equally old (or older) Denisovan individual are closer to their common ancestral population than present-day populations. This explanation is compatible with the observation that the numbers of substitutions inferred to have occurred between the Denisovan genome sequence and the common ancestor of humans and chimpanzees is smaller than for present-day humans (15, 16).

1. Kircher M, Sawyer S, Meyer M (2012) Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform. *Nucleic Acids Res* 40(1):e3.
2. Kircher M, Stenzel U, Kelso J (2009) Improved base calling for the Illumina Genome Analyzer using machine learning strategies. *Genome Biol* 10(8):R83.
3. Kircher M, Heyn P, Kelso J (2011) Addressing challenges in the production and analysis of Illumina sequencing data. *BMC Genomics* 12:382.
4. Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25(14):1754–1760.
5. Li H, et al.; 1000 Genome Project Data Processing Subgroup (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25(16):2078–2079.
6. Briggs AW, et al. (2009) Targeted retrieval and analysis of five Neandertal mtDNA genomes. *Science* 325(5938):318–321.
7. DeAngelis MM, Wang DG, Hawkins TL (1995) Solid-phase reversible immobilization for the isolation of PCR products. *Nucleic Acids Res* 23(22):4742–4743.
8. Hodges E, et al. (2007) Genome-wide in situ exon capture for selective resequencing. *Nat Genet* 39(12):1522–1527.
9. Patterson N, et al. (2012) Ancient admixture in human history. *Genetics* 192(3): 1065–1093.
10. Gnirke A, et al. (2009) Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. *Nat Biotechnol* 27(2):182–189.
11. Kircher M (2012) Analysis of high-throughput ancient DNA sequencing data. *Methods Mol Biol* 840:197–228.
12. Green RE, et al. (2008) A complete Neandertal mitochondrial genome sequence determined by high-throughput sequencing. *Cell* 134(3):416–426.
13. Reich D, et al. (2010) Genetic history of an archaic hominin group from Denisova Cave in Siberia. *Nature* 468(7327):1053–1060.
14. Reich D, et al. (2011) Denisova admixture and the first modern human dispersals into Southeast Asia and Oceania. *Am J Hum Genet* 89(4):516–528.
15. Skoglund P, Jakobsson M (2011) Archaic human ancestry in East Asia. *Proc Natl Acad Sci USA* 108(45):18301–18306.
16. Meyer M, et al. (2012) A high-coverage genome sequence from an archaic Denisovan individual. *Science* 338(6104):222–226.

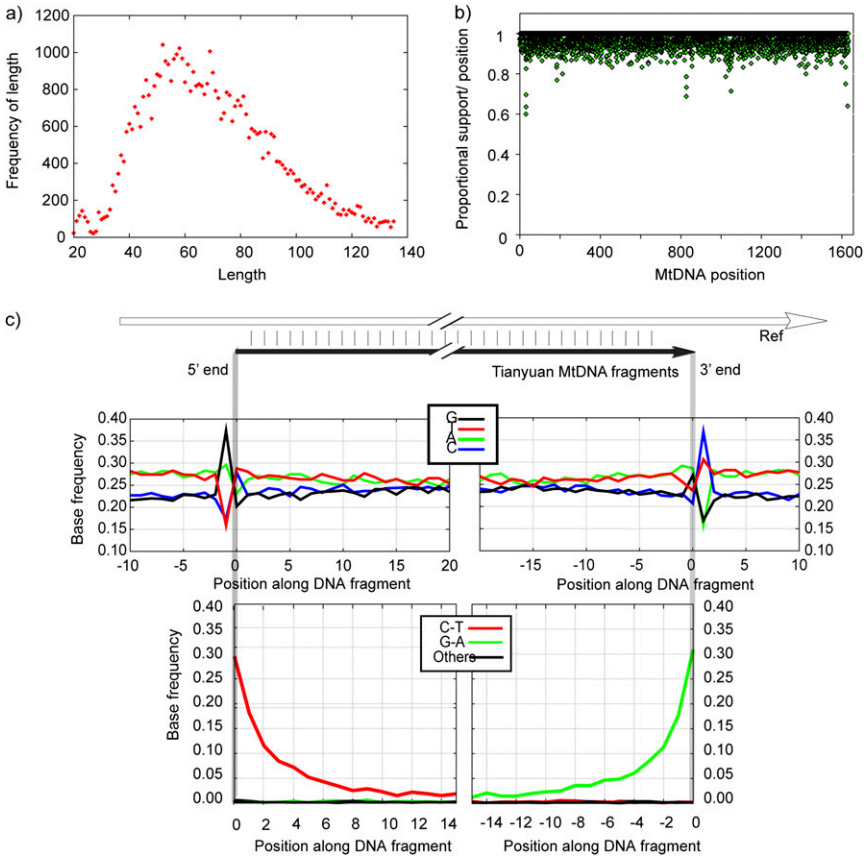

Fig. S1. (*A*) Length distribution of Tianyuan mtDNA fragments. (*B*) Fraction of bases agreeing with the consensus sequence along the mitochondrial genome. (*C*) DNA fragmentation and nucleotide misincorporation patterns inferred from Tianyuan mtDNA fragments. The base composition of the reference genome around fragment ends (Upper) indicates preferential strand breakage 3 of the purines. Increased frequency of C→T and G→A substitutions (Lower) close to fragment ends indicates cytosine deamination. The frequency X→Y is given as the fraction of bases at the indicated positions along the DNA fragments and inferred to be X according to the consensus sequence, which is observed as Y in the individual DNA fragments.

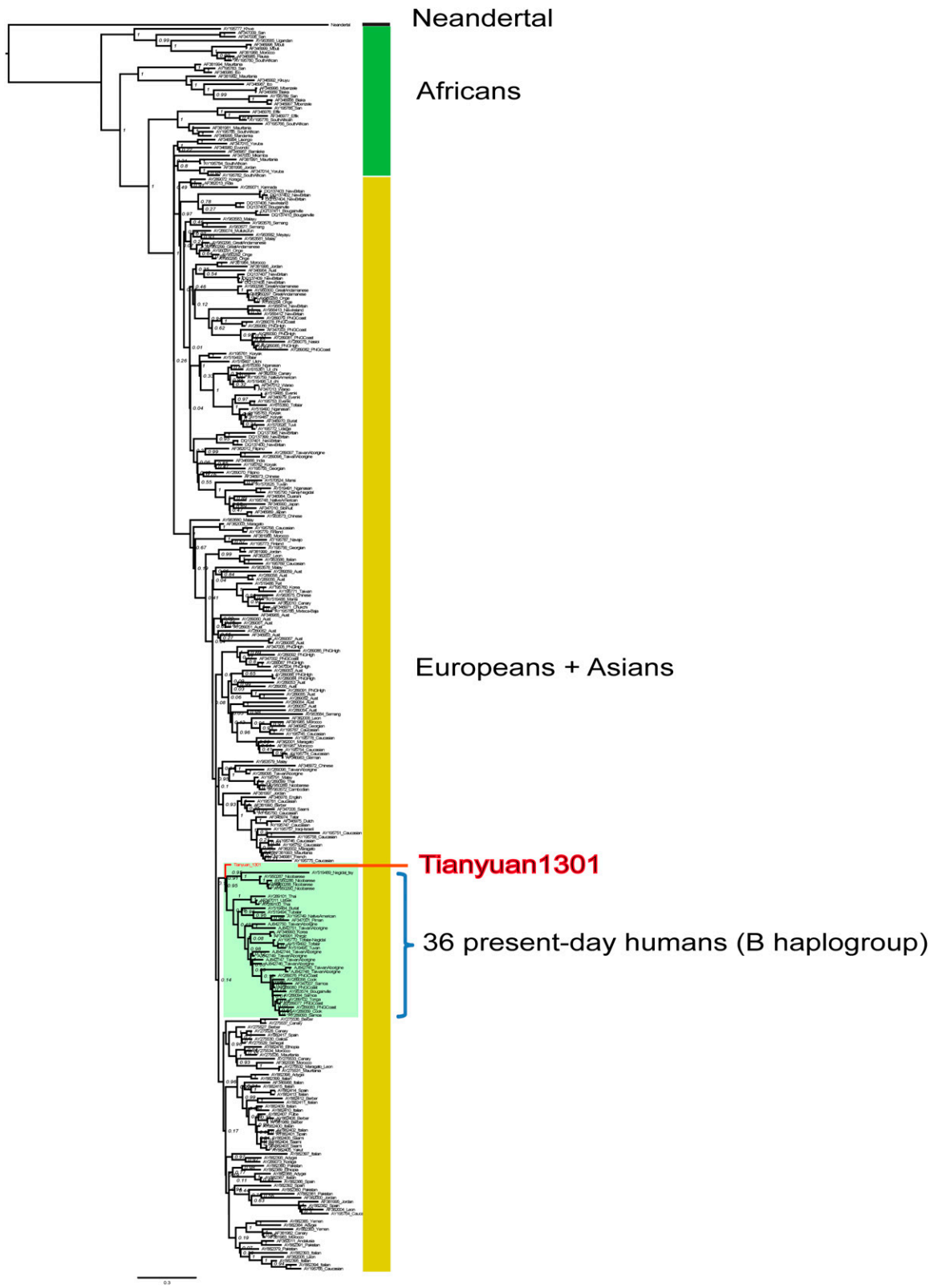

Fig. S2. Phylogenetic tree of mtDNAs from Tianyuan and 311 present-day humans. The phylogeny was estimated with a Bayesian approach under a GTR+I model of sequence evolution, using a Neandertal mtDNA (Vindija 33.25) (1) as an outgroup. Bar colors indicate different populations. The clade of Tianyuan and its closest 36 present-day humans is in green.

1. Briggs AW, et al. (2009) Targeted retrieval and analysis of five Neandertal mtDNA genomes. *Science* 325(5938):318–321.

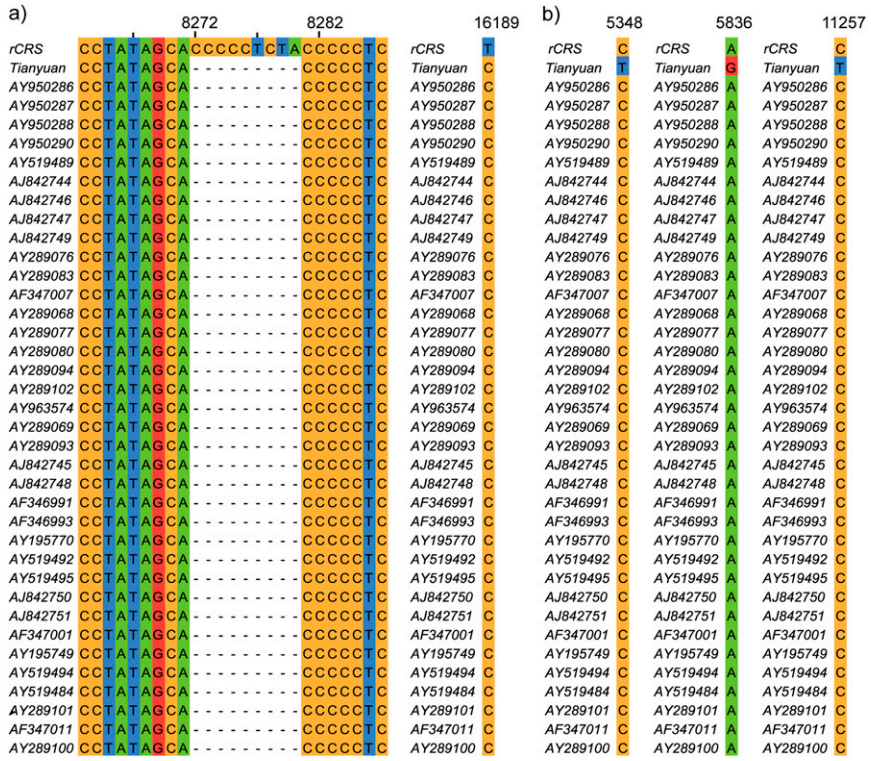

Fig. S3. (*A*) A 9-bp deletion and a substitution (at position 16,189) present in Tianyuan mtDNA and present-day human mtDNAs belonging to haplogroup B. (*B*) Three positions where at least 308 of 311 present-day human mtDNAs agree and the Tianyuan mtDNA is different.

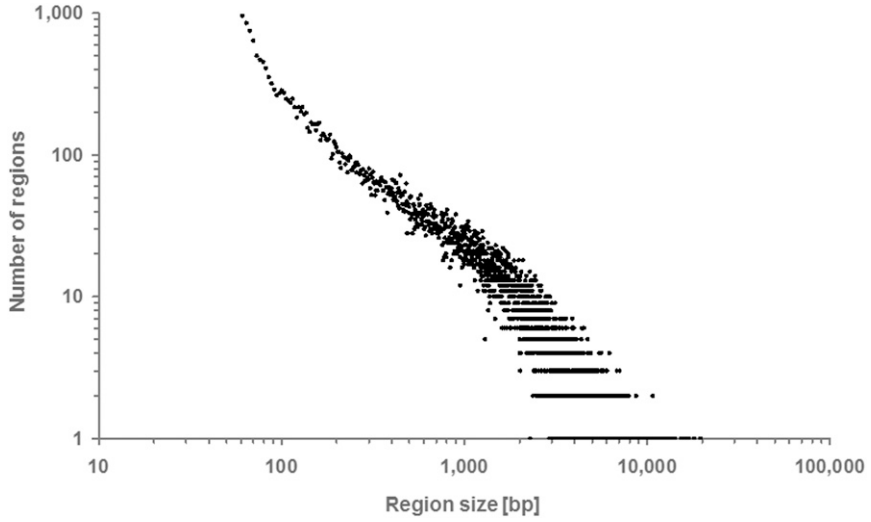

Fig. S4. Size distribution of chromosome 21 target regions.

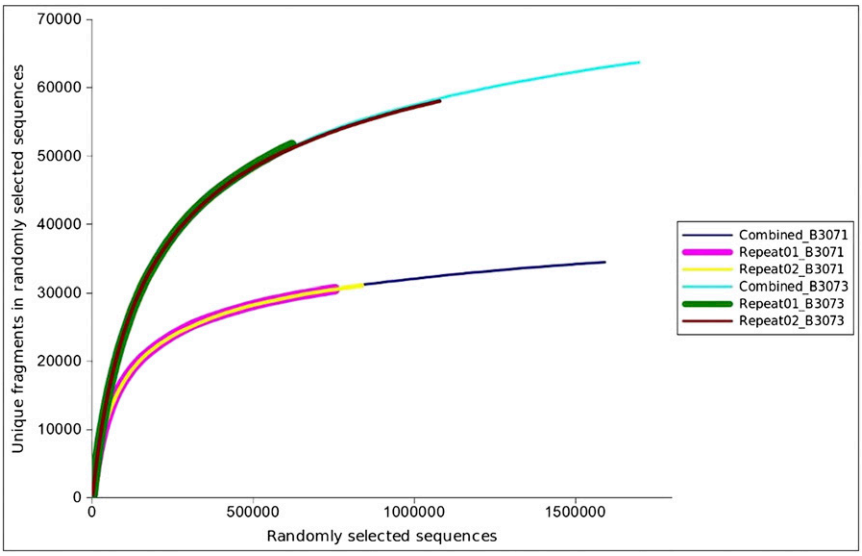

Fig. S5. Subsampling plots obtained from sequences from replicate captures of chromosome 21. Plots are shown for two of the libraries only (B3071 and B3073), but the results are consistent with those seen for all libraries.

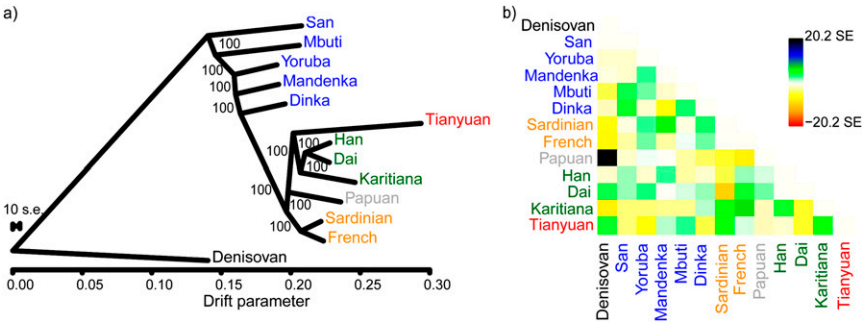

Fig. S6. (*A*) Maximum-likelihood tree of chromosome 21 sequences of the Tianyuan individual, 11 present-day humans, and the Denisovan individual. (*B*) Residual matrix of chromosome 21 sequences of the Tianyuan individual, 11 present-day humans, and the Denisovan individual. Residuals above zero represent populations that are closely related to each other according to the allele covariance matrix. SE indicates the average SE across each pair of populations. Darker colors represent stronger signal.

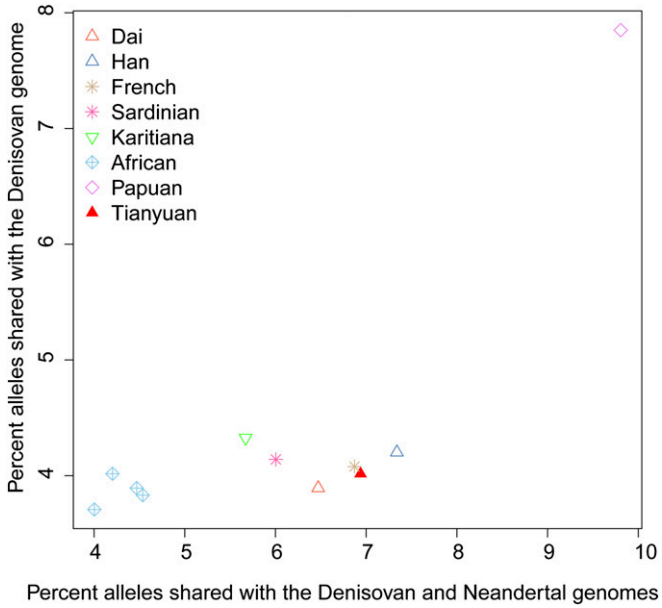

Fig. S7. Proportions of alleles shared with Neandertal and Denisovan genomes in the Tianyuan and 11 present-day individuals. The *x* axis shows the percent of alleles that match the Neandertal and Denisovan genomes at sites where both of these differ from the seven Africans, and the *y* axis indicates the percent of alleles that match the Denisovan genome where this differs from the Neandertal as well as the seven Africans.

Table S1. Shotgun and mtDNA sequencing information for TY1301 and TY1305

Bone ID	Library ID	Powder (mg)	Unique fragments aligned to hg19	Endogenous (%)	Average length (bp)	Unique fragments aligned to mtDNA	Coverage of mtDNA	Coverage of mtDNA/mg
TY1301-1	B3071	37	2,529	0.02	64	4,949	20	0.54
TY1301-2	B3073	84	2,993	0.03	52	9,883	35.6	0.42
TY1305-1	B3072	61	2,263	0.01	59	653	2.4	0.04
TY1305-2	B3074	93	725	0.01	42	625	1.7	0.02

Table S2. Distance matrix between all possible pairs of genotypes

	AA	AB	BB	A	B
AA	0	1	2	0	2
AB	1	0	1	1	1
BB	2	1	0	2	0
A	0	1	2	0	2
B	2	1	0	2	0

A, reference allele; B, alternative allele; double letters indicate diploid calls; single letters indicate haploid calls.