Long-term genetic stability and a high-altitude East Asian origin for the peoples of the high valleys of the Himalayan arc

Choongwon Jeonga, Andrew T. Ozgaa, David B. WUUAAWUWskyny, Helena Malmstrømb, Hanna Edlundd, Courtney A. Hofmab, Richard W. Haganb, Matthias Jakobssonb, Cecil M. Lewisb, Mark S. Aldenderfer,1 Anna Di Rienzoa,1 and Christina Warinnerb,1,e

aDepartment of Human Genetics, University of Chicago, Chicago, IL 60637; bDepartment of Anthropology, University of Oklahoma, Norman, OK 73019; cDepartment of Organismal Biology, Uppsala University, Uppsala SE-752 36, Sweden; dSchool of Social Sciences, Humanities, and Arts, University of California, Merced, CA 95343; and Institute of Evolutionary Medicine, University of Zurich, Zurich CH-8057, Switzerland

Edited by Lynn B. Jorde, University of Utah, Salt Lake City, UT, and accepted by Editorial Board Member Richard G. Klein May 17, 2016 (received for review October 21, 2015)

The high-altitude transverse valleys (≥3,000 m above sea level (masl)) of the Himalayan arc from Arunachal Pradesh to Ladakh were among the last habitable places permanently colonized by prehistoric humans due to the challenges of resource scarcity, cold stress, and hypoxia. The modern populations of these valleys, who share cultural and linguistic affinities with peoples found today on the Tibetan plateau, are commonly assumed to be the descendants of the earliest inhabitants of the Himalayan arc. However, this assumption has been challenged by archaeological and osteological evidence suggesting that these valleys may have been originally populated from areas other than the Tibetan plateau, including those at low elevation. To investigate the peopling and early population history of this dynamic high-altitude contact zone, we sequenced the genomes (0.04×–7.25×, mean 2.16×) and mitochondrial genomes (20.8×–1,311.0×, mean 482.1×) of eight individuals dating to three periods with distinct material culture in the Annapurna Conservation Area (ACA) of Nepal, spanning 3,150–1,250 y before present (yBP). We demonstrate that the region is characterized by long-term stability of the population genetic make-up despite marked changes in material culture. The ancient genomes, uniparental haplotypes, and high-altitude adaptive alleles suggest a high-altitude East Asian origin for prehistoric Himalayan populations.

Significance

Since prehistory, the Himalayan mountain range has presented a formidable barrier to population migration, whereas at the same time its transverse valleys have long served as conduits for trade and exchange. Yet, despite the economic and cultural importance of Himalayan trade routes, little is known about the region’s peopling and early population history. In this study, we conduct to our knowledge the first ancient DNA investigation of the Himalayan arc and generate genome data for eight individuals ranging in time from the earliest known human settlements to the establishment of the Tibetan Empire. We demonstrate that the region was colonized by East Asians of likely high-altitude origin, followed by millennia of genetic continuity despite marked changes in material culture and mortuary behavior.


This article is a PNAS Direct Submission. L.B.J. is a guest editor invited by the Editorial Board. Freely available online through the PNAS open access option.

Data deposition: Metagenomic DNA sequences have been deposited in the NCBI Short Read Archive (SRA) (project accession no. SRP065070 and sample accession nos. SRR2751054–SRR2751058, SRR2751060–SRR2751063, SRR2751066–SRR2751067, SRR2751070, SRR2751142, SRR2751148, SRR2751152, SRR3222643, SRR3222649, SRR3222655, SRR3222659, SRR3222661, SRR3222664, SRR3222686, SRR3222749, SRR3222758, SRR3222765, and SRR3222772).

1To whom correspondence may be addressed. Email: christina.warinner@ou.edu, maldenderfer@merced.edu, or dirienzo@bsd.uchicago.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1520844113/-/DCSupplemental.

www.pnas.org/cgi/doi/10.1073/pnas.1520844113

PNAS Early Edition | 1 of 6
Foodstuffs within ACA prehistoric funerary contexts include domesticates of both West Asian (e.g., barley, buckwheat, lentils, peas, sheep, and goats) and East Asian (e.g., rice) origin (24). In addition to locally made utilitarian wares, prestige objects include copper ornaments and vessels, carnelian beads, marine shell pendants, and faience suggestive of a strong South Asian connection, as well as bamboo baskets, mats, and cups and wooden furniture and design motifs suggesting contact with Central Asia and Xinjiang (12, 25). Later periods after ca. 1,750 yBP before present (yBP) also include Chinese silk and glass beads from Sassanida (modern-day Iran) and central and far southern India, as well as gold and silver masks that resemble those found in western Tibet, Ladakh, and Kyrgyzstan (14). Finally, local mortuary practices initially resemble those observed in northern Xinjiang, but after ca. 1,500 yBP include defleshing, a practice that may have multiple origins but is primarily associated with Western Asian cultures (23). Therefore, there is evidence that early populations in the Himalayan transverse valleys were exposed to influences from a remarkably wide geographic extent, from Iran to eastern China.

Given the complexity in material culture, currently available archaeological data cannot determine whether population replacement, cultural diffusion, or both are responsible for these diverse influences. Furthermore, interpretation of linguistic and genetic data from present-day populations is complicated by multiple historically documented Tibetan migrations, after ca. 1,300 yBP, linked to the rise and fall of the Tibetan Empire, extensive warfare, and the establishment of modern nation states (26, 27). For these reasons, the analysis of ancient human genomes provides a unique and direct means for resolving competing hypotheses regarding the population history of the high Himalayas.

To investigate the peopling and early population history of the ACA, we obtained genome-wide sequences and high-coverage mitochondrial sequences from eight individuals from three periods with distinct material culture: Chokhopani (3,150–2,400 yBP), Mebrak (2,400–1,850 yBP), and Samdzong (1,750–1,250 yBP) (Table 1). Following initial population affinity analyses, we then further sequenced the genomes of five individuals to >2× coverage to obtain higher-resolution genome data and increase the coverage of two genes associated with high-altitude adaptation, EPAS1 (endothelial PAS domain protein 1) and EGLN1 (e.g.l family hypoxia-inducible factor 1) and EPAS1 (endothelial PAS domain protein 1). Our results are consistent with long-term genetic stability in the region; additionally, genome sequences, uniparental haplotypes, and high-altitude adaptive alleles support a high-altitude East Asian origin for these prehistoric Himalayan populations.

Results
Ancient DNA Extraction and Sequencing Quality. Eight prehistoric ACA dental samples (C1, M63, M240, M344, S10, S35, S40, and S41) were sequenced in the first phase of this study and found to contain relatively high proportions of human DNA, ranging from 2.6% to 58.3% (SI Appendix, Table S1). Five of these samples were selected for deeper sequencing. This included three Samdzong period samples (S10, S35, and S41) containing >40% human reads and the oldest sample in the study, C1, dating to the Chokhopani period, containing 31.0% human reads, each of which was sequenced to >2× mean coverage. A Mebrak period sample, M63, with 18.9% human reads was also sequenced to 1× mean coverage. In total, mean sequence coverage at a genome-wide level for all eight samples ranged between 0.04× and 7.25×, and between 20.8× and 1,311× for the mitochondrial genome (SI Appendix, Table S1). Genetic sex was confidently assigned for all eight individuals, of which seven were male (SI Appendix, Fig. S1). Given the comparatively low proportions of human DNA reported in previous ancient DNA (aDNA) studies, the preservation of the ACA samples is very good, which is consistent with the arid and cold burial environment and relatively low thermal age of the sites (28).

Assessment of Contamination from Modern Humans. After initial alignment, we assessed whether the human reads we recovered were likely to be endogenous (i.e., not resulting from modern contamination) by examining chemical damage patterns typical of aDNA (29–31) and estimating the proportion of contaminant reads from mtDNA sequences (32). We observed typical ancient DNA damage patterns in all of the ACA samples, suggesting that the vast majority of DNA is of ancient origin. First, human DNA sequences were short in length, with median lengths of 55–87 bp (SI Appendix, Table S1 and Fig. S2). Second, 8.8–19.0% of sequences exhibited terminal 5′–C > T miscoding lesions (SI Appendix, Fig. S3), a characteristic pattern of aDNA damage. Finally, purines (A and G) were enriched at 5′–1 positions (SI Appendix, Fig. S4), indicating depurination-driven strand breaks, another characteristic pattern of aDNA damage.

These features qualitatively support a high proportion of endogenous DNA in the ACA samples. However, the dataset can still contain a small number of contaminant human reads. Therefore, we estimated the proportion of contaminant mitochondrial reads, using a Bayesian method implemented in the program contamMix (32). The estimated proportion of endogenous reads in the ACA samples is >98% for all samples except M344 (94.4%), suggesting minimal contamination from other humans (SI Appendix, Table S1).

Genome-Wide SNP Profiling of Ancient DNA Samples. To understand the genetic relationship between the ACA aDNA samples and populations around the world, we compared sequences from our first-phase sequencing data to genetic data of 26 contemporary populations from the 1,000 genomes (1KG) project and high-coverage (>30×) Illumina-sequenced whole genomes of 17 modern humans, including 4 Sherpa and 2 Tibetans from Nepal. Overlapping each aDNA sample dataset with the above population genetic data panel, we retrieved 0.47–6.36 million autosomal SNPs for our first-phase samples (SI Appendix, Table S1). All eight ACA individuals across the three time periods were found to be most closely related to East Asians (SI Appendix, Figs. S5–S8), a finding consistently supported by the results of several approaches, including principal components analysis (PCA), model-based unsupervised genetic clustering, and the outgroup f3 statistic. The latter is a measure of genetic affinity that

Table 1. ACA dental samples investigated in this study

<table>
<thead>
<tr>
<th>Period/site</th>
<th>Dates</th>
<th>No. samples</th>
<th>Sample ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chokhopani</td>
<td>3,150–2,400 yBP</td>
<td>1</td>
<td>C1</td>
</tr>
<tr>
<td>Mebrak</td>
<td>2,400–1,850 yBP</td>
<td>3</td>
<td>M63, M240, M344</td>
</tr>
<tr>
<td>Samdzong</td>
<td>1,750–1,250 yBP</td>
<td>4</td>
<td>S10, S35, S40, S41</td>
</tr>
</tbody>
</table>

Fig. 1. Map of the ACA and sampling locations. The ACA (dark gray), located in the Upper Mustang of north-central Nepal and bordering Tibet (inset), is situated between the Annapurna and Dhaulagiri Massifs of the main Himalayan mountain range. The ACA includes 14 mountains in excess of 6,000 masl, and it contains a single major drainage, the Kali Gandaki River, which originates on the Tibetan plateau. Data from ref. 69.
measures the branch length from an outgroup to the split point of a pair of populations (33).

To refine our inferences of genetic affinity, we further sequenced five ACA individuals to 1.0–7.3x coverage and compared the resulting genotypes to array genotyping data from Tibetans (34), Sherpas (18), and populations from the Human Genome Diversity Panel (35), as well as whole-genome sequences of two Nepali Tibetans. Multiple lines of evidence consistently indicate high-altitude East Asians (i.e., the Sherpa and Tibetans) as the closest contemporary populations to the ACA individuals, regardless of time period. First, ACA individuals cluster together with Tibetans in PCA (Fig. 2). Second, model-based unsupervised clustering inferences that a large proportion of ancestry in the ACA individuals is shared with the Sherpa and Tibetans (Fig. 3). Third, all ACA individuals have the largest outgroup statistic with the Sherpa and Tibetans, followed by other Tibeto-Burman–speaking groups such as Naxi, Yi, and Tuja (Fig. 4A and SI Appendix, Figs. S8 and S9). Finally, formal comparison of population affinity in the form of the D test shows that all of the ACA individuals are more closely related to Tibetans from Lhasa (Z = 2.7–8.0 SD), Tibetans from Nepal (Z = 0.8–4.6 SD), and the Sherpa (Z = 2.5–6.8 SD) than to any other population (Fig. 4B and SI Appendix, Fig. S10).

Additionally, outgroup statistic (SI Appendix, Fig. S12A) and D tests (SI Appendix, Fig. S12B) support, albeit slightly less consistently, a greater genetic affinity of contemporary high-altitude populations with the ACA samples than with the Yi or Naxi (S41 is an exception, possibly due to a minor west-Eurasian component) (Fig. 3; SI Appendix; and SI Appendix, Fig. S12). Taken as a whole, our results strongly suggest that the ACA individuals are closely related to contemporary high-altitude East Asian populations.

High-Altitude Functional Alleles. Encouraged by the genetic profiles of the ACA individuals, we investigated whether the five more deeply sequenced ACA individuals share high-altitude adaptive genetic variants (15–17) with Tibetan populations (SI Appendix, Table S2). More specifically, we determined whether they have derived alleles at 20 noncoding SNPs that tag the selected haplotype at the EPAS1 gene (36) or at two nonsynonymous SNPs (rs12097901 and rs186996510) with signatures of adaptive allele frequency divergence at the EGLN1 gene (22, 37). Currently, there is broad agreement for selection on the derived EGLN1 alleles beginning ca. 8,000 ya (22, 37), but dating the onset of selection for the derived EPAS1 haplotype has proved more controversial. The derived EPAS1 haplotype was recently shown to have originated in the Denisova genome and its presence in the human genome represents a recent archaic introgression (36). Consequently, the subsequent selection of this haplotype in humans is difficult to model using genetic data from living populations, and dates ranging from 2,750 ya to 18,250 ya have been proposed (13, 17, 38).

Interestingly, all reads from our ACA individuals match the derived allele for the nonsynonymous EGLN1 SNP rs186996510 (SI Appendix, Table S2), including the oldest Chokhophani sample (C1). This derived allele, c.12G > C (p.Asp4Glu), is reported in high frequency in Tibetans (0.64–0.85) (22, 37), but is rare in low-altitude East Asians (0.03 in 1KG phase 3 East Asians) and virtually absent outside East Asia. Functional studies have implicated this allele as playing a role in oxygen homeostasis under hypoxic conditions (37, 39). In contrast, reads supporting derived alleles at the EPAS1 SNPs were found in two of the three later Samdong individuals (S35 and S41), but not in the earlier Chokhophani (C1) or Mebrak (M63) individuals. This observation of shared adaptive alleles between ancient ACA individuals and contemporary Tibetans is consistent with our genome sequence results suggesting that the ACA inhabitants are affiliated with contemporary high-altitude East Asians. In addition, the contrasting pattern of alleles for the two genes leads us to speculate that the EGLN1 and EPAS1 adaptive haplotypes rose to high frequency at different time points in these ancient high-altitude populations, although more samples must be sequenced to accurately estimate allele frequency change across time.

Mitochondrial and Y Chromosome Haplogroup Identification. Using high-coverage, consensus full mtDNA genome sequences (SI Appendix, Table S1), we next inferred haplogroup assignment for each ACA individual. All eight individuals are assigned to haplogroups reported to be present in contemporary Nepalis and/or Tibetans (SI Appendix, Table S3) (21, 40) and rare or absent in present-day Indian and Pakistani populations (41). The oldest sample in our study, C1, belongs to haplogroup D4, a major maternal lineage among Tibetans. Interestingly, Tibetan D4 has a deep divergence time from other East Asian populations (26–27 kya), further supporting genetic affinity between the ACA individuals and contemporary high-altitude East Asians (21). Four male individuals with >2x coverage were
much more gradual in the north than in the south. Thus, ascending populations on the north side may have been able to stay at intermediate altitudes for extended periods of time, allowing for acclimatization and the accumulation of genetic and subsistence adaptations, whereas potential migrants from the south side had no access to such a buffer zone because of the limited availability of sufficient habitable land at intermediate altitudes. This scenario is supported by the archaeological record of the Tibetan plateau. Archaeological data from the northeastern Tibetan plateau indicate an initial occupation ca. 15,000 ya (20, 46), long before the colonization of the high-traverse valleys in the Himalayan arc. Archaeological data also support later influences from the East Asian side of the plateau associated with the appearance of agriculture after 5,500 yBP, evidenced by the adoption of Neolithic domesticates, first from East Asia (millets and pigs) and later from West Asia (via Central Asia: barley, sheep, and goats). It has been proposed that these changes enabled populations on the plateau to move to higher and more marginal lands after ca. 4,000 yBP (47), where they may have subsequently served as a source population for the Himalayan transverse valleys. It is beyond the scope of our current study, however, to address whether the spread of agriculture onto the plateau was accompanied by population migration.

Genetic adaptation to high altitude also likely facilitated this asymmetric colonization. Accumulation of beneficial mutations is a feature expected for a population gradually adapting to a new environment. Evolution of such beneficial mutations across time provides crucial information for understanding the strength and cause of natural selection. Contemporary high-altitude East Asians on the Tibetan plateau have at least two such genes, EPASI and EGLN1, that exhibit strong signatures of positive natural selection as well as functional properties consistent with an adaptive role in high-altitude environments (15, 16, 37, 39, 48). Importantly, we found that the oldest Chokhopani sample (C1) and three later Samdzong individuals (S10, S35, and S41) are most likely homozygous for a derived nonsynonymous allele of the EGLN1 SNP (Ex3:696C > T; S10), suggesting that this allele was already segregating in the prehistoric founding population. In contrast, derived alleles from the EPASI SNPs were observed only in Samdzong individuals, implying an asynchronous evolution of the two genes. However, sequencing of additional ancient samples through time is necessary to reconstruct the adaptive evolution of these and other beneficial mutations in the ACA. Given the unusually high quality of aDNA from the ACA, population-level ancient genome sequencing is likely an achievable goal once additional early archaeological specimens are available.

It is tempting to compare this case to archaeogenetic studies in Europe, which suggest that large-scale cultural transitions are frequently associated with massive population movements (49–52). In the Himalayas, we observe two discrete cultural transitions (associated with the Mebrak and Samdzong periods) without evidence of changes in the genetic makeup of the population. One sample from Samdzong (S41) may be an exception in that it is the only one showing some amount of non-East Asian ancestry; however, this proportion is estimated to be small (Fig. 3). Therefore, the predominance of East Asian ancestry in the ACA samples supports our hypothesis that certain topographies, specifically very high altitudes, require a unique set of adaptations, genetic or cultural, that differ from those sufficient for low-altitude migration and colonization. However, because current archaeological data are largely limited to funerary contexts, we caution that the archaeological changes we observe in the ACA may not represent full-scale cultural transitions.

In this study, we conducted to our knowledge the first successful ancient DNA investigation of prehistoric Himalayan populations and retrieved high proportions of endogenous aDNA from eight high-altitude ACA individuals dating to three distinct cultural periods spanning 3,150–1,250 yBP. Our population genetic analysis strongly supports the genetic affiliation of prehistoric Himalayan populations with contemporary East Asians and at a subcontinental level suggests a closer affinity with present-day high-altitude East Asians, such as Tibetans and Sherpa, than with low-altitude East Asians. Moreover, this affinity is consistent through time, suggesting
that temporal changes in material culture and mortuary behavior largely reflect acculturation or cultural diffusion rather than large-scale gene flow or population replacement from outside East Asia. Finally, we provide to our knowledge the first empirical evidence for differing evolutionary dynamics of selection on the EGLNI and EPAS1 genes in prehistoric high-altitude populations. Considering the pivotal role of the Himalayan high transverse valleys in connecting far-flung Eurasian populations, as well as the environmental challenges they impose on their inhabitants, our study has deep implications for the understanding of human migration history and adaptation to local environments and for future genetic archaeology studies.

Experimental Procedures

Study Design and Samples. The ACA of Upper Mustang, Nepal is located in northern central Nepal and covers an area of ~7,630 km² (Fig. 1 and SI Appendix, section 1). Prior archaeological research in the region identified three distinct periods of occupation: Chokhopani (3,150–2,400 yBP) (45, 53), Mebrak (2,400–1,850 yBP) (12, 53), and Samdzong (1,750–1,250 yBP) (14, 23), each defined by a type site of the same name. Dental samples from 12 individuals were selected for DNA screening (SI Appendix, Table S3), of which 8 yielded sufficient data for continental-level ancestry analysis (Table 1). Of these, 3 were more deeply sequenced to investigate questions regarding regional ancestry and high-altitude adaptation. Use of ancient and preexisting, deidentified modern human genetic data was determined to be exempt from human subjects review (University of Chicago IRB12-1785).

Ancient DNA Extraction, Library Construction, and Sequencing. DNA extraction was performed in a dedicated ancient DNA facility in accordance with established contamination control precautions and workflows, as previously described (28) (SI Appendix, section 2). Following decalcification and digestion, two DNA extraction methods were compared: (i) phenol-chloroform separation followed by purification and concentration using a MiniElute PCR Purification kit (Qiagen) (28) and (ii) salting out followed by purification and concentration using a QiAamp DNA Mini Kit (54). For 3 of the 12 individuals, DNA extraction was performed with both methods. Purified DNA was quantified using a Qubit High Sensitivity dsDNA assay (Life Technologies). DNA extracts were built into indexed illumina libraries, using a double-stranded library protocol, with minor modifications (SI Appendix, section 3). The resulting libraries were purified, quantified, and pooled for sequencing on the illumina HiSeq platform, using paired-end 100-bp, 125-bp, or 150-bp chemistry (SI Appendix, Table S6).

Sequence Data Filtering and Quality Control. Adapter sequences were removed and each read pair was merged into a single sequence, using a publicly available python script (55) (https://bioinf.eva.mpg.de/fastqProcessing/MergeReadsWithQc.py). Merged reads were mapped to the human reference genome hg19, using BWA-backtrack 0.7.9a (56). Uniquely mapped reads ≥35 bp were kept, and PCR duplicates were removed, keeping the one with the highest mapping quality score. Step-by-step filtering and quality parameters and statistics are provided in SI Appendix, section 4 and Table S1.

Comparison of DNA Extraction Methods. Before proceeding further, the performance of the two DNA extraction methods was compared (SI Appendix, section 5). The phenol-chloroform/MiniElute method substantially outperformed the salting out/QIAamp method in both total DNA yield and human DNA content (SI Appendix, Table S5 and Fig. S11). Consequently, all subsequent genetic analyses were restricted to the eight samples (C1, M240, M3, S6, S10, S35, S40, and S41) extracted using the phenol-chloroform/MiniElute method.

Assessment of Genetic Sex, Sample Contamination, and DNA Damage. Genetic sex was estimated using previously described methods for shotgun sequence data (57) (SI Appendix, section 6). Contamination was assessed by estimating the proportion of endogenous reads among human mitochondrial DNA sequences, using the Bayesian program contaminX (32) (SI Appendix, section 7). For each sample, the estimated endogenous content and 95% confidence interval are provided in SI Appendix, Table S1. DNA fragment lengths and damage patterns typical of ancient DNA were assessed from uniquely mapped, nonduplicate reads, using the mapDamage program (30, 31) (SI Appendix, section 8).

Data Filtering and Compilation for Population Genetic Analysis. For population genetic analysis, we retrieved ACA genetic information from sequence reads (SI Appendix, section 9). High-quality base calls (≥Q30) from reads with high mapping quality scores (≥30) were collected for each genomic position, using the mpileup command of SAMTools v1.2 (56), after masking 5 bp at both ends of reads to reduce the effect of cytosine deamination. For the analysis of the first-phase data (<1x coverage), one read at each position was then randomly sampled to generate haploid genotypes. ACA aDNA data were then overlapped with available genetic variation data of 26 worldwide populations from the 1KG project phase 3 haplotype set (ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502), which includes 2,504 unrelated individuals, and high-coverage (≥30x) published (18, 58, 59) and unpublished (https://www.simonsfoundation.org/life-sciences/simons-genome-diversity-project-dataset/) whole-genome sequences of 17 modern humans from 10 populations, including Tibetans and Sherpa. Finally, high-coverage genotypes from the chimpanzee genome assembly Pan_troglobyes-2.1.4 (panTro4), Altai Neandertal (59), and Denisovan (38) were compiled to assess ancestral alleles and archaic hominin ancestry. Variants within human repeat regions or CpG islands (60), variable sites with multiple alternative alleles, sites with strand ambiguity (A/T or G/C SNPs), sites prone to cytosine deamination, and sites not present in the 1KG dataset were removed for PCA, clustering, and outgroup f3 analyses. Additionally, sites with missing genotypes among the 17 modern humans or nonhuman samples were also excluded. This process led to variable numbers of SNPs for each aDNA sample, ranging from 0.47 million to 6.36 million SNPs. For the analysis of the second-phase data (1.0–7.3x coverage), we sampled a read for ~650,000 positions in the Human Genome Diversity Panel (HGDP) (35). Additional array genotyping data for the Sherpa (18) and Tibetans (34) were intersected with the sequence data of two contemporary Tibetan individuals. For genetic clustering, we used genotype likelihoods calculated from the GATK v2.7-4 UnifiedGenotyper module (61) (SI Appendix, section 9).

Whole-Genome Ancestry Affiliation Analysis. PCA was performed using the smartpca program in the EIGENSOFT 6.0 package (62). For analysis of the first-phase data, PCA was run for each ancient sample separately, using 0.15–2.03 million SNPs with minor allele frequency (MAF) ≥ 0.01. Results for PC1 and PC2 were then merged by a Procrustes transformation, using the “procrustes” function in the R package “shaper” (63) (SI Appendix, section 10). For analysis of the second-phase data, the “gssoptject: YES” option was used to project ancient samples onto the PC plane, calculated with 357,000 SNPs with MAF ≥ 0.01. We performed model-based genetic clustering analysis as implemented in the snMF (64) and NSGadmix (65) programs. For analysis of the first-phase data, one allele from each modern sample was randomly sampled at each variable site to match the haploid nature of the aDNA samples. SNPs with minor allele frequency (MAF) ≥ 0.01, with minor allele frequency (MAF) ≥ 0.01, and high-coverage genotypes from the chimpanzee genome assembly Pan_troglobyes-2.1.4 (panTro4), Altai Neandertal (59), and Denisovan (38) were compiled to assess ancestral alleles and archaic hominin ancestry. Variants within human repeat regions or CpG islands (60), variable sites with multiple alternative alleles, sites with strand ambiguity (A/T or G/C SNPs), sites prone to cytosine deamination, and sites not present in the 1KG dataset were removed for PCA, clustering, and outgroup f3 analyses. Additionally, sites with missing genotypes among the 17 modern humans or nonhuman samples were also excluded. This process led to variable numbers of SNPs for each aDNA sample, ranging from 0.47 million to 6.36 million SNPs. For the analysis of the second-phase data (1.0–7.3x coverage), we sampled a read for ~650,000 positions in the Human Genome Diversity Panel (HGDP) (35). Additional array genotyping data for the Sherpa (18) and Tibetans (34) were intersected with the sequence data of two contemporary Tibetan individuals. For genetic clustering, we used genotype likelihoods calculated from the GATK v2.7-4 UnifiedGenotyper module (61) (SI Appendix, section 9).

High-Altitude Adaptation Allele Analysis. For sites with a read depth ≥1, allelic variants were determined in the EGLNTI gene and in 20 tagging SNPs in the EPAS1 gene (SI Appendix, section 11).

Uniparental Haplogroup Analysis. Consensus mtDNA sequences for all eight individuals were called from sequence reads, using the UnifiedGenotype module of the GATK v2.7-4 followed by haplogroup assignment using Haplogrep2 on individual aDNA reads. For analysis of the second-phase data, 105,944 LD-pruned SNPs with MAF > 0.01 were used, and 50 replicates were performed for K values of 2–9. Genetic affinity was estimated using the outgroup-certainty (33) from the KRG Yoruba (YRI) or HGDP Yoruba as an outgroup and using the D statistic, using the snMF for PCA. The qp3Pop and qpDstat programs in the ADMIXTools v2 package (67) were used to calculate f3 and D statistics and associated SEs (SI Appendix, section 10). In addition to comparison with ACA samples, f3 and D statistics were also calculated comparing high-altitude East Asian and lowland Tibetan-Burman-speaking populations (SI Appendix, Fig. S12).

ACKNOWLEDGMENTS. We thank the Department of Archaeology of the Government of Nepal for permission to conduct archaeological research in the ACA and for access to curated samples, Pete Athans for mountaineering assistance, Jacqueline Eng for osteological assistance, Raul Tito for assistance with sample preparation, Arielle Reivant Munters for technical assistance, John Novembre for helpful discussions during data analysis, and the local residents of the ACA who have assisted our research team during many seasons of fieldwork. Computational resources for data analysis were provided by the Beagle supercomputer at the Computation Institute and by the Center for Research Informatics at the University of Chicago. This work was supported by the US National Science Foundation [Grant ATMHSD-057620 (to M.S.A.) and Grant BCS-1528698 (to M.S.A., C.W., and A.D.R.)], the US National Institutes of

Jeong et al. | PNAS Early Edition | 5 of 6


