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

Choongwon Jeong[a], Andrew T. Ozga[b], David B. Witonsky[a], Helena Malmström[c], Hanna Edlund[c], Courtney A. Hofman[b], Richard W. Hagan[b], Mattias Jakobsson[c], Cecil M. Lewis[b], Mark S. Aldenderfer[d,1], Anna Di Rienzo[a,1], and Christina Warinner[b,e,1]

[a]Department of Human Genetics, University of Chicago, Chicago, IL 60637; [b]Department of Anthropology, University of Oklahoma, Norman, OK 73019; [c]Department of Organismal Biology, Uppsala University, Uppsala SE-752 36, Sweden; [d]School of Social Sciences, Humanities, and Arts, University of California, Merced, CA 95343; and [e]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 Ladahk 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.

Ancient DNA | population genetics | high altitude | EGLN1 | EPAS1

The world's high plateaus and great mountain ranges were among the last places colonized by humans in prehistory (1–3). The challenges of rough terrain, cold stress, hypoxia, and the relative scarcity of resources in these high places significantly slowed the pace at which permanent occupation took place. The Himalayan mountain range and the Tibetan plateau are among the highest places on earth. The Himalayas include 9 of the 10 tallest mountains in the world, and, at an average elevation of 5,000 m above sea level (masl), the Tibetan Plateau is ∼25% higher than the Peruvian altiplano, the next highest plateau in the world (1, 4, 5). Genome-wide studies of the geographic structure of modern populations clearly point to the Himalayas as a barrier to gene flow between East Asians and Western Eurasians (6). However, there is also extensive evidence of cultural and linguistic diversity across the Himalayan arc, which hints at a long history of cross-regional contact (7, 8).

Currently available archaeological and osteological data and genetic data from modern-day populations have been used to support contrasting hypotheses of South Asian (9–11), Central Asian (12), lowland Southeast Asian (13), and high-altitude East Asian (7, 8) origins for the earliest Himalayan inhabitants, and there is likewise little agreement regarding subsequent regional population history (12, 14). Successful permanent habitation of high-altitude environments requires numerous physiological adaptations, and recent genetic studies have identified robust signals of positive natural selection underlying adaptations to hypoxia in Tibetans (15–17) and in the Sherpa (18), an ethnic group that migrated from the eastern Tibetan plateau to Nepal 400–600 y ago (ya) (19). Tibetans and Sherpa are the only two present-day high-altitude East Asian ethnic groups that have been studied to date, using genome-wide markers. Although the Tibetan plateau has been the subject of intense study regarding its population history and high-altitude adaptation (15, 20–22), far less is known about the much later colonization of the surrounding high transverse valleys along the Himalayan arc. Elucidating this history is important because these valleys have long served as natural corridors and trade routes connecting the Tibetan plateau to the Indian subcontinent. Moreover, the role of adaptation to high-altitude hypoxia in the initial colonization of these valleys and in the subsequent gene flow through them is entirely unexplored.

The Annapurna Conservation Area (ACA) of Upper Mustang, Nepal (Fig. 1) is a major high-elevation corridor (2,800–4,500 masl) that includes the earliest known archaeological sites containing preserved human remains in a Himalayan transverse valley (23).

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.

Author contributions: C.M.L., M.S.A., A.D.R., and C.W. designed research; A.T.O., H.M., H.E., C.A.H., R.W.H., M.S.A., and C.W. performed research; M.J. and A.D.R. contributed new reagents/analytic tools; C.J., D.B.W., and C.W. analyzed data; and C.J. and C.W. wrote the paper.

The authors declare no conflict of interest.

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. SRR2751055–SRR2751058, SRR2751060–SRR2751063, SRR2751066–SRR2751067, SRR2751070, SRR2751142, SRR2751148, SRR2751152, SRR3222643, SRR3222649, SRR3222655, SRR3222659, SRR3222661, SRR3222664, SRR3222686, SRR3222749, SRR3222758, SRR3222765, and SRR3222772).

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

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

Downloaded at Akademiska Sjukhuset on December 20, 2019

ANTHROPOLOGY

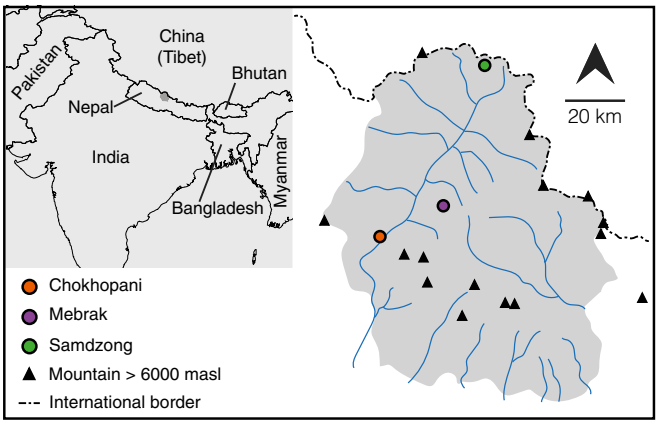

Fig. 1. Map of the ACA and sampling locations. The ACA (dark gray), located in the Upper Mustang of north-central Nepal and bordering Tibet (*Inset*), is situated between the Annapurna and Dhaulagiri Massifs of the main Himalayan mountain range. The ACA includes 14 mountains in excess of 6,000 masl,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

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 y before present (yBP) also include Chinese silk and glass beads from Sassania (modern-day Iran) and central and far southern India, as well as gold and silver masks that resemble those found in western Tibet, Ladahk, 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 dating to 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, *EGLN1* (egl-9 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 0.044× and 7.253× and between 20.8× and 1,311.0× 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 analyses (*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 f_3 statistic. The latter is a measure of genetic affinity that

Table 1. ACA dental samples investigated in this study

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

Downloaded at Akademiska Sjukhuset on December 20, 2019

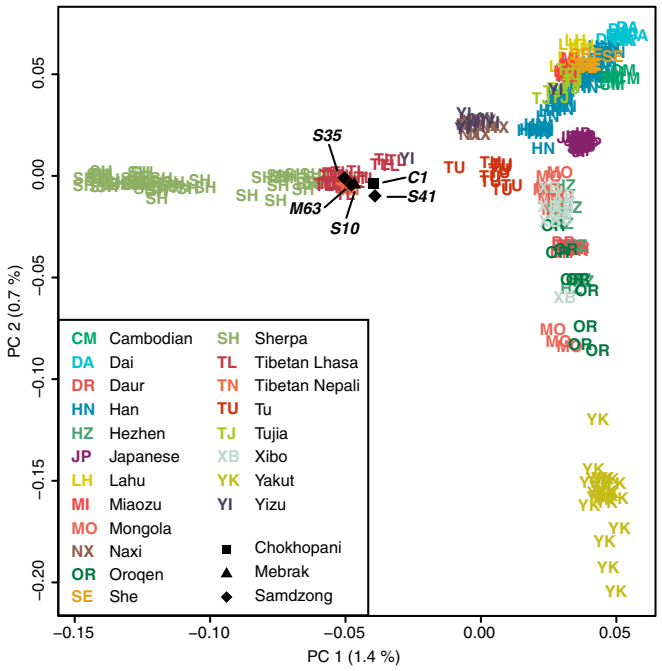

Fig. 2. PCA of East Asian populations and ancient ACA individuals. All five ACA samples cluster with Tibetans. PC1 and PC2 were calculated using all contemporary East Asian samples. Ancient ACA samples were projected onto the PC plane, using the "*lsqproject: YES*" option.

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 1.0–7.3× coverage and compared the resulting genotypes to array genotyping data from Tibetans (34), Sherpa (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 infers 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 f_3 statistic with the Sherpa and Tibetans, followed by other Tibeto-Burman–speaking groups such as Naxi, Yi, and Tujia (Fig. 4*A* 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. 4*B* and *SI Appendix*, Fig. S10).

Additionally, outgroup f_3 (*SI Appendix*, Fig. S12*A*) and *D* tests (*SI Appendix*, Fig. S12*B*) 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 Chokhopani 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 Samdzong individuals (S35 and S41), but not in the earlier Chokhopani (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 >2× coverage were

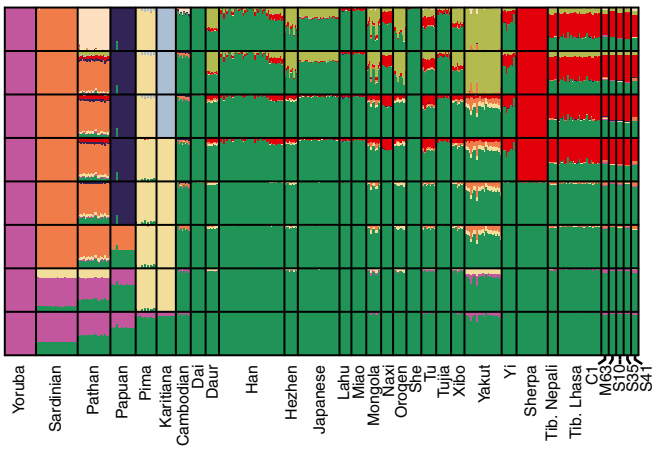

Fig. 3. Unsupervised genetic clustering with two to nine ancestral populations (K = 2–9). All five ACA samples exhibit ancestry profiles most similar to Tibetans. The Sherpa, Tibetans, and ACA samples share a distinct high-altitude ancestry (red) in the highest proportions, followed by other Tibeto-Burman–speaking groups such as Naxi and Yi. K = 2 is shown at the bottom of the plot; K = 9 is shown at the top.

ANTHROPOLOGY

Downloaded at Akademiska Sjukhuset on December 20, 2019

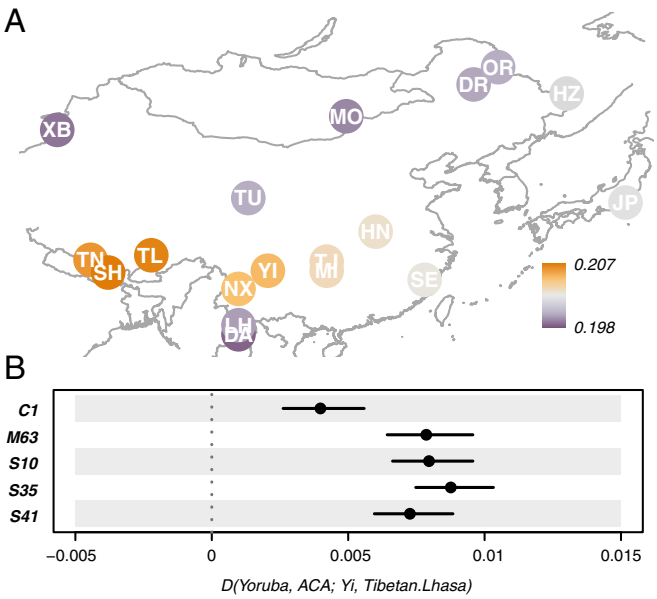

Fig. 4. Genetic affinity of ACA individuals and East Asian populations, using genome-wide SNP data. (*A*) Genetic affinity with ancient sample C1 is measured by f_3 (Yoruba; ACA, X). For all ACA samples (*SI Appendix*, Fig. S9), either Sherpa or Tibetans were the closest modern population (a larger f_3 value indicates a closer relationship), followed by other Tibeto-Burman–speaking groups, such as Naxi and Yi. (*B*) All ACA samples are significantly more closely related to contemporary high-altitude East Asians than they are to lowland Tibeto-Burman–speaking groups, as shown by positive values of Patterson's *D* (Yoruba, ACA; Yi, Tibetan.Lhasa). Equivalent results are observed if the test is performed with alternative proxy populations for high-altitude East Asians (e.g., Tibetan.Nepali or Sherpa) and lowland Tibeto-Burman speakers (e.g., Naxi or Tujia) (*SI Appendix*, Fig. S10).

determined to belong to Y chromosome haplogroups O-M117 and D (*SI Appendix*, Table S4), which are among the most frequent haplogroups in contemporary Tibetans (21) based on haplogroup-tagging SNPs from the cleantree program (42).

Discussion

The role of geography in migration and population structure has been a central topic in population genetics studies of our species and others (43, 44). At a genetic level, the Himalayan arc delineates a sharp genetic barrier between South Asian and East Asian populations, a striking anomaly against a general isolation-by-distance pattern of human population structure across much of Eurasia (6). The asymmetric topography of the Himalayan massif, bordered by a high-elevation plateau to the north and lowland plains to the south, is reflected in the current regional genetic structure of human populations, evidenced by autosomal and Y chromosome STR (short tandem repeat) frequencies in modern Nepalese populations (7, 8). Our results suggest that the Himalayas have long served as a remarkably resistant barrier to northward but not southward gene flow and that this genetic boundary has been stable for at least the last three millennia. The eight ancient ACA individuals analyzed in this study exhibit a strong and consistent genetic affiliation to contemporary East Asian, and especially to high-altitude East Asian (Sherpa and Tibetan), populations. Therefore, previous proposals of a South Asian (9–11), Central Asian (12), or low-elevation Southeast Asian (13) origin of the first inhabitants of this region, as well as speculation regarding subsequent prehistoric population replacement or large-scale admixture with lowland populations (12, 14, 45), are not supported.

It is interesting that the high-altitude barrier to migration seems to be more permeable from the northern, as opposed to the southern, side of the Himalayan arc. One can speculate that this disparity may be due to the topographical differences between the northern and southern sides of the arc: the altitudinal gradient is 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. Archeological 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, *EPAS1* 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 (rs186996510), suggesting that this allele was already common in the founding population. In contrast, derived alleles from the *EPAS1* 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

Downloaded at Akademiska Sjukhuset on December 20, 2019

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 *EGLN1* 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 km^2 (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 S5), of which 8 yielded sufficient data for continental-level ancestry analysis (Table 1). Of these, 5 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 MinElute 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 using 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 platforms, 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/MergeReadsFastQ_cc.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/MinElute 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, M344, M63, S10, S35, S40, and S41) extracted using the phenol-chloroform/MinElute 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 contamMix (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 (<1× 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 (≥ 30×) 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_troglodytes-2.1.4 (panTro4), Altai Neandertal (59), and Denisovan (58) 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 f_3 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 (1.0–7.3× 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 HGDP dataset, as well as genotypes of two contemporary Nepali 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) ≥ 0.001. Results for PC1 and PC2 were then merged by a Procrustes transformation, using the "procGPA" function in the R package "shapes" (63) (*SI Appendix*, section 10). For analysis of the second-phase data, the "*lsqproject: 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 NGSadmix (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 ACA aDNA samples. SNPs with at least three copies of the minor allele were retained, and linkage disequilibrium (LD)-based SNP pruning was performed using PLINK v1.0.7 (66) ($r^2 > 0.2$). The resulting SNPs were downsampled to 40,000–100,000 to match the number of SNPs available in each ACA sample, and analysis was performed in 50 replicates with random seeds for two to eight clusters (*K*) (*SI Appendix*, section 10). 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 f_3 statistic (33) with 1KG Yoruba (YRI) or HGDP Yoruba as an outgroup and using the *D* statistic, using the SNP set for PCA. The *qp3Pop* and *qpDstat* programs in the ADMIXtools v2 package (67) were used to calculate f_3 and *D* statistics and associated SEs (*SI Appendix*, section 10). In addition to comparison with ACA samples, f_3 and *D* statistics were also calculated comparing high-altitude East Asian and lowland Tibeto-Burman–speaking populations (*SI Appendix*, Fig. S12).

High-Altitude Adaptation Allele Analysis. For sites with a read depth ≥1, allelic variants were determined in the *EGLN1* 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 HaploGrep (68) (*SI Appendix*, section 12). For five male individuals analyzed in the second phase of the study, the Y haplogroup was manually assigned based on reads piled up for 519 informative biallelic SNPs from a database associated with cleantree software (*SI Appendix*, section 13).

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 ATM/HSD-057620 (to M.S.A.) and Grant BCS-1528698 (to M.S.A., C.W., and A.D.R.)], the US National Institutes of Health

Downloaded at Akademiska Sjukhuset on December 20, 2019

[Award R01HL119577 (to A.D.R.)], the European Research Council [Starting Grant 311413 (to M.J.)], the University of Chicago Comprehensive Cancer Center Support Grant P30 CA14599 (to A.D.R.) with particular support from the Genomics Core Facility, the National Geographic Society (M.S.A.), the Henry Luce Foundation (M.S.A.), a Samsung Scholarship (to C.J.), The North Face company (Pete Athans), and the Field Museum.

1. Aldenderfer M (2006) Modelling plateau peoples. *World Archaeol* 38(3):357–370.
2. Bellwood PS (2014) *The Global Prehistory of Human Migration* (Wiley, Hoboken, NJ).
3. Gamble C (1994) *Timewalkers* (Harvard Univ Press, Cambridge, MA).
4. Harris N (2006) The elevation history of the Tibetan Plateau and its implications for the Asian monsoon. *Palaeogeogr Palaeocl* 241(1):4–15.
5. Houston J, Hartley AJ (2003) The central Andean west-slope rainshadow and its potential contribution to the origin of hyper-aridity in the Atacama desert. *Int J Climatol* 23(12):1453–1464.
6. Wang C, Zöllner S, Rosenberg NA (2012) A quantitative comparison of the similarity between genes and geography in worldwide human populations. *PLoS Genet* 8(8):e1002886.
7. Gayden T, et al. (2009) Genetic insights into the origins of Tibeto-Burman populations in the Himalayas. *J Hum Genet* 54(4):216–223.
8. Gayden T, et al. (2013) The Himalayas: Barrier and conduit for gene flow. *Am J Phys Anthropol* 151(2):169–182.
9. Stacul G (1969) Excavations near Ghaligai and chronological sequence of protohistorical cultures in the Swat Valley. *East and West* 19(1-2):44–91.
10. Singh A (1999) Cist burials in Kinnaur, Western Himalayas. *Cent Asiatic J* 43(2): 249–258.
11. Hüttel H (1997) Archäologische Siedlungsforschung im Hohen Himalaya: Die Ausgrabungen de KAVA im Muktinath Tal/Nepal 1994-1995 [Research on archaeological settlements in the high Himalayas: The KAVA excavations in the Muktinath Valley, Nepal 1994-1995]. *Beitr Allgem Vergleich Archäol* 17:7–64. German.
12. Alt KW, et al. (2003) Climbing into the past - first Himalayan mummies discovered in Nepal. *J Archaeol Sci* 30(11):1529–1535.
13. Peng MS, et al. (2011) Inland post-glacial dispersal in East Asia revealed by mitochondrial haplogroup M9a'b. *BMC Biol* 9:2.
14. Aldenderfer M (2013) Variation in mortuary practice on the early Tibetan plateau and the high Himalayas. *J Int Assoc Bon Res* 1:293–318.
15. Beall CM, et al. (2010) Natural selection on EPAS1 (HIF2alpha) associated with low hemoglobin concentration in Tibetan highlanders. *Proc Natl Acad Sci USA* 107(25):11459–11464.
16. Simonson TS, et al. (2010) Genetic evidence for high-altitude adaptation in Tibet. *Science* 329(5987):72–75.
17. Yi X, et al. (2010) Sequencing of 50 human exomes reveals adaptation to high altitude. *Science* 329(5987):75–78.
18. Jeong C, et al. (2014) Admixture facilitates genetic adaptations to high altitude in Tibet. *Nat Commun* 5:3281.
19. Oppitz M (1974) Myths and facts: Reconsidering some data concerning the clan history of the Sherpas. *Kailash* 2(1-2):121–131.
20. Aldenderfer M (2011) Peopling the Tibetan plateau: Insights from archaeology. *High Alt Med Biol* 12(2):141–147.
21. Qi X, et al. (2013) Genetic evidence of paleolithic colonization and neolithic expansion of modern humans on the Tibetan plateau. *Mol Biol Evol* 30(8):1761–1778.
22. Xiang K, et al. (2013) Identification of a Tibetan-specific mutation in the hypoxic gene EGLN1 and its contribution to high-altitude adaptation. *Mol Biol Evol* 30(8):1889–1898.
23. Aldenderfer M, Eng J (2016) Death and burial at two ancient high altitude communities of Nepal. *The Companion to South Asia in the Past*, eds Schug G, Walimbe S (Wiley-Blackwell, London).
24. Knorzer KH (2000) 3000 years of agriculture in a valley of the High Himalayas. *Veg Hist Archaeobot* 9(4):219–222.
25. Massa G (2013) The funerary metals of Samdzong. MSc thesis (University College London, London).
26. La Polla R (2001) The role of migration and language contact in the development of the Sino-Tibetan language family. *Areal Diffusion and Genetic Inheritance*, eds Aikhenvald AY, Dixon RMW (Oxford Univ Press, Oxford), pp 225–254.
27. Childs G (2012) Trans-Himalayan migrations as processes, not events. *Origins and Migrations in the Extended Eastern Himalaya*, eds Huber T, Blackburn S (Brill, Leiden, The Netherlands), pp 11–32.
28. Ziesemer KA, et al. (2015) Intrinsic challenges in ancient microbiome reconstruction using 16S rRNA gene amplification. *Sci Rep* 5:16498.
29. Briggs AW, et al. (2007) Patterns of damage in genomic DNA sequences from a Neandertal. *Proc Natl Acad Sci USA* 104(37):14616–14621.
30. Ginolhac A, Rasmussen M, Gilbert MT, Willerslev E, Orlando L (2011) mapDamage: Testing for damage patterns in ancient DNA sequences. *Bioinformatics* 27(15):2153–2155.
31. Jónsson H, Ginolhac A, Schubert M, Johnson PL, Orlando L (2013) mapDamage2.0: Fast approximate Bayesian estimates of ancient DNA damage parameters. *Bioinformatics* 29(13):1682–1684.
32. Fu Q, et al. (2013) A revised timescale for human evolution based on ancient mitochondrial genomes. *Curr Biol* 23(7):553–559.
33. Raghavan M, et al. (2014) Upper Palaeolithic Siberian genome reveals dual ancestry of Native Americans. *Nature* 505(7481):87–91.
34. Wang B, et al. (2011) On the origin of Tibetans and their genetic basis in adapting high-altitude environments. *PLoS One* 6(2):e17002.
35. Li JZ, et al. (2008) Worldwide human relationships inferred from genome-wide patterns of variation. *Science* 319(5866):1100–1104.
36. Huerta-Sánchez E, et al. (2014) Altitude adaptation in Tibetans caused by introgression of Denisovan-like DNA. *Nature* 512(7513):194–197.
37. Lorenzo FR, et al. (2014) A genetic mechanism for Tibetan high-altitude adaptation. *Nat Genet* 46(9):951–956.
38. Hackinger S, et al. (2016) Wide distribution and altitude correlation of an archaic high-altitude-adaptive EPAS1 haplotype in the Himalayas. *Hum Genet* 135(4):393–402.
39. Petousi N, et al. (2014) Tibetans living at sea level have a hyporesponsive hypoxia-inducible factor system and blunted physiological responses to hypoxia. *J Appl Physiol* 116(7): 893–904.
40. Fornarino S, et al. (2009) Mitochondrial and Y-chromosome diversity of the Tharus (Nepal): A reservoir of genetic variation. *BMC Evol Biol* 9:154.
41. Metspalu M, et al. (2004) Most of the extant mtDNA boundaries in south and southwest Asia were likely shaped during the initial settlement of Eurasia by anatomically modern humans. *BMC Genet* 5:26.
42. Ralf A, van Oven M, Zhong K, Kayser M (2015) Simultaneous analysis of hundreds of Y-chromosomal SNPs for high-resolution paternal lineage classification using targeted semiconductor sequencing. *Hum Mutat* 36(1):151–159.
43. Novembre J, et al. (2008) Genes mirror geography within Europe. *Nature* 456(7218):98–101.
44. Petkova D, Novembre J, Stephens M (2016) Visualizing spatial population structure with estimated effective migration surfaces. *Nat Genet* 48(1):94–100.
45. Tiwari D (1984/85) Cave burials from western Nepal, Mustang. *Ancient Nepal* 85:1–12.
46. Zhao M, et al. (2009) Mitochondrial genome evidence reveals successful Late Paleolithic settlement on the Tibetan Plateau. *Proc Natl Acad Sci USA* 106(50):21230–21235.
47. Guedes Jd A, et al. (2014) Moving agriculture onto the Tibetan Plateau. *Archaeol Anthropol Sci* 6(3):255–269.
48. Yu-jing S, et al. (2010) Endothelial nitric oxide synthase gene polymorphisms associated with susceptibility to high altitude pulmonary edema in Chinese railway construction workers at Qinghai-Tibet over 4 500 meters above sea level. *Chin Med Sci J* 25(4):215–221.
49. Allentoft ME, et al. (2015) Population genomics of Bronze Age Eurasia. *Nature* 522(7555):167–172.
50. Haak W, et al. (2015) Massive migration from the steppe was a source for Indo-European languages in Europe. *Nature* 522(7555):207–211.
51. Skoglund P, et al. (2012) Origins and genetic legacy of Neolithic farmers and hunter-gatherers in Europe. *Science* 336(6080):466–469.
52. Skoglund P, et al. (2014) Genomic diversity and admixture differs for Stone-Age Scandinavian foragers and farmers. *Science* 344(6185):747–750.
53. Simons A, Schön W (1998) Cave systems and terrace settlements in Mustang, Nepal: Settlement periods from prehistoric times up to the present day. *Beitr Allgem Vergleich Archäol* 18:27–47.
54. Tito RY, et al. (2011) Brief communication: DNA from early Holocene American dog. *Am J Phys Anthropol* 145(4):653–657.
55. Kircher M (2012) Analysis of high-throughput ancient DNA sequencing data. *Ancient DNA: Methods and Protocols*, eds Shapiro BA, Hofreiter M (Springer, New York).
56. Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25(14):1754–1760.
57. Skoglund P, Stora J, Gotherstrom A, Jakobsson M (2013) Accurate sex identification of ancient human remains using DNA shotgun sequencing. *J Archaeol Sci* 40(12):4477–4482.
58. Meyer M, et al. (2012) A high-coverage genome sequence from an archaic Denisovan individual. *Science* 338(6104):222–226.
59. Prüfer K, et al. (2014) The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature* 505(7481):43–49.
60. Wu H, Caffo B, Jaffee HA, Irizarry RA, Feinberg AP (2010) Redefining CpG islands using hidden Markov models. *Biostatistics* 11(3):499–514.
61. McKenna A, et al. (2010) The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* 20(9):1297–1303.
62. Patterson N, Price AL, Reich D (2006) Population structure and eigenanalysis. *PLoS Genet* 2(12):e190.
63. Dryden I (2014) Shapes package. Contributed package Version 1.1-10. R Foundation for Statistical Computing, Vienna.
64. Frichot E, Mathieu F, Trouillon T, Bouchard G, François O (2014) Fast and efficient estimation of individual ancestry coefficients. *Genetics* 196(4):973–983.
65. Skotte L, Korneliussen TS, Albrechtsen A (2013) Estimating individual admixture proportions from next generation sequencing data. *Genetics* 195(3):693–702.
66. Purcell S, et al. (2007) PLINK: A tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 81(3):559–575.
67. Patterson N, et al. (2012) Ancient admixture in human history. *Genetics* 192(3): 1065–1093.
68. Kloss-Brandstätter A, et al. (2011) HaploGrep: A fast and reliable algorithm for automatic classification of mitochondrial DNA haplogroups. *Hum Mutat* 32(1):25–32.
69. Banskota K, Sharma B (1995) *Tourism for Mountain Community Development: Case Study Report on the Annapurna and Gorkha Regions of Nepal* (International Centre for Integrated Mountain Development, Kathmandu, Nepal).

Downloaded at Akademiska Sjukhuset on December 20, 2019