Limited South Asian admixture in ethnic Tibetans from the Himalayan valleys in Nepal

The southern slope of the Himalayas is a zone of contact and admixture between various populations of South and East Asian origins [1, 2]. Although previous studies found that recent genetic contribution from South Asian gene pools rapidly decreases when moving from low to high altitudes [1, 2], presence and magnitude of South Asian ancestry in the ethnic Tibetan groups we studied here have not been studied so far. To characterize South Asian admixture in the Himalayan Tibetans in our study, we first compiled a genome-wide genotype data of world-wide populations from published studies, either array-genotyped on Illumina platforms [3-11] or whole-genome sequenced [5, 7, 8, 12-28]. We intersected SNPs between different platforms and filtered low-quality SNPs and individuals (SNPs and individuals with > 5% missing rate per data set) using PLINK v1.9 [29], resulting in 3,256 individuals and 224,268 autosomal SNPs, including our new data of 991 individuals (S9 Table).

We first performed a principal component analysis (PCA) of Eurasian populations using the smartpca v16000 program in the EIGENSOFT v7.2 package [30]. Because PCA is sensitive to sample size and our Himalayan Tibetans have much bigger sample sizes than the rest of groups in the panel, we calculated PCs without Himalayan Tibetan individuals; instead, we projected them onto PCs calculated from a published panel of 1,295 Eurasian individuals using the “lsqproject: YES” option (S9 Table). This approach also provides an unbiased way of comparing related individuals in PCA. We repeated the same analysis using 357 East Asian and Siberian individuals with no detectable Western Eurasian ancestry (S9 Table).

Eurasian PCA results show that all of the Himalayan Tibetans form a cluster close to other East Asians but far from South Asians (S7 Fig). In the East Asian PCA, they fall on next to present-day Tibetans from near Lhasa (S8 Fig). These results show that there is no large-scale South Asian admixture in the Himalayan Tibetans residing at 3,000 m or higher in altitude.

We calculated outgroup-f3 statistic of the form $f_3(Mbuti; Himalayan Tibetan, world-wide)$, a measure of shared genetic drift between Himalayan Tibetans and world-wide populations since their split from a common outgroup, Central African rainforest hunter-gatherer Mbuti, for each Himalayan Tibetan group (Nubri, Tsum, Lower Mustang and Upper Mustang). For all four Himalayan groups, Tibetans from Lhasa (“Tibetan”) [11] and Sherpas from the Khumbu region (“Sherpa”) [10] show the highest outgroup-f3 values (S9 Fig). We further tested if the Himalayan Tibetans have extra allele sharing with South Asians by calculating $f_4(Mbuti, South Asian; Tibetan/Sherpa, Himalayan)$. Nubri and Lower Mustang show significantly positive $f_4$ statistics ($f_4 \geq 3$ standard errors, SEs; S9 Fig). The other two, Tsum and Upper Mustang, show positive statistics but within statistical noise (< 3 SEs). Using the qpAdm program, a generalization of $f_4$ statistic that tests symmetry between the target and the admixture model against
multiple outgroups [20], we find that Sherpa + Pathan provides a good fit to the Himalayan Tibetans with 2.8-6.2% South Asian contribution \( (\chi^2 \ p \geq 0.310; \text{S10 Table}) \). In line with significant positive \( f_4 \) statistic with South Asians and other Western Eurasians, the Himalayan Tibetans cannot be modeled as a sister group of Sherpa without the above estimated small amount of South Asian contribution \( (\chi^2 \ p \leq 4.22 \times 10^{-12}) \). We used qp3Pop (v400), qpDstat (v711) and qpAdm (v632) programs from the ADMIXTOOLS v3.0 package [31]. For the qpAdm analysis, we used eight outgroups: Mbuti (n=13), Onge (n=2), Dai (n=10), Chukchi (n=12), Nganasan (n=12), Neolithic Anatolian farmers (“Anatolia_N”; n=23) [20], Western European Mesolithic hunter-gatherers (“WHG”; n=3) [15, 21] and Eneolithic Yamnaya individuals from Samara (“Yamnaya_Samara”; n=9) [22].

We applied the same qpAdm model for each individual separately to estimate South Asian admixture proportion without being affected by sample size or population-specific drift. Per-individual South Asian ancestry proportion ranges 0% to 11.6%, with the mean of each group matching with group-based estimates (S10 Fig). South Asian ancestry proportion is negatively correlated with altitude even after taking into account sub-region labels \( (p = 2.89 \times 10^{-9}; \text{S11 Fig}) \).

Admixture dates are estimated to be between 17 and 31 generations ago (505-892 years before present, using 29 years per generation; S12 Fig) using weighted admixture LD-based method implemented in ALDER v1.3 [32]. We used Sherpas and Pathans as two references and minimum distance of 1.0 cM for fitting the exponential curve.

References

