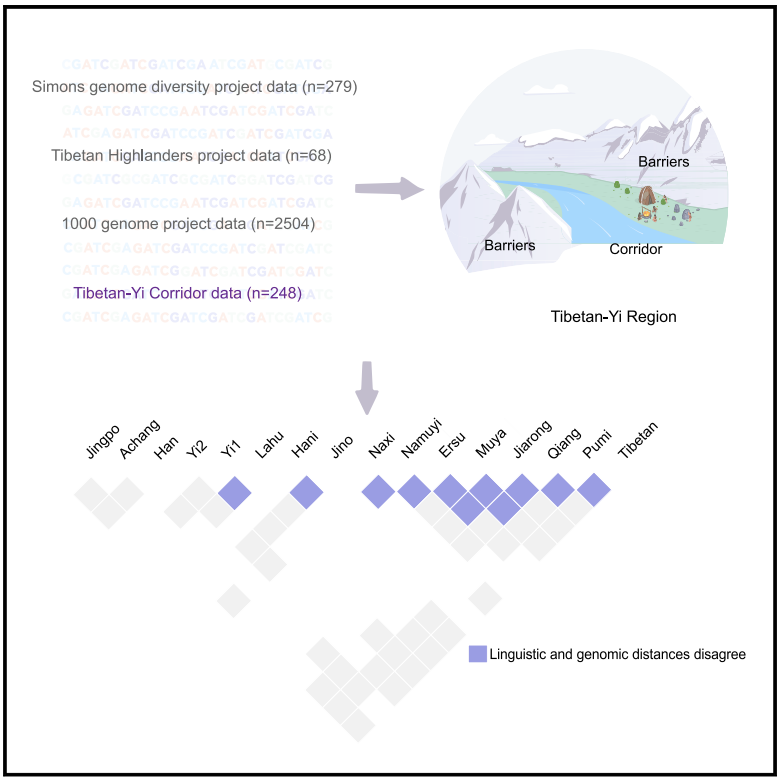
Cell Reports

Article

The Tibetan-Yi region is both a corridor and a barrier for human gene flow

Graphical abstract



Highlights

- TYC populations inferred by linguistics are incompatible with the genetic evidence
- The slopes of the Tibetan plateau form a barrier to gene flow
- Genetic variation largely follows geographic patterns
- The Achang shows evidence of prolonged isolation and genetic drift

Authors

Zhe Zhang, Yanlin Zhang, Yinan Wang, ..., Rasmus Nielsen, Shuai Cheng Li, Shengbin Li

Correspondence

rasmus_nielsen@berkeley.edu (R.N.), shuaicli@cityu.edu.hk (S.C.L.), shengbinlee@mail.xjtu.edu.cn (S.L.)

In brief

The Tibetan-Yi Corridor is a crossroads between northern and southern east Asia. Zhang et al. report the genomic data of 248 minority individuals in this region to reconstruct the pattern of population movement and differentiation between diverse Asian populations.

Zhang et al., 2022, Cell Reports *39*, 110720
April 26, 2022
https://doi.org/10.1016/j.celrep.2022.110720

CellPress

Cell Reports

CellPress
OPEN ACCESS

Article

The Tibetan-Yi region is both a corridor and a barrier for human gene flow

Zhe Zhang,[1,9,10] Yanlin Zhang,[8,10] Yinan Wang,[1,10] Zicheng Zhao,[8,10] Melinda Yang,[3] Lin Zhang,[4] Bin Zhou,[4] Bingying Xu,[5] Hongbo Zhang,[1] Teng Chen,[1] Wenkui Dai,[2] Yong Zhou,[2] Shuo Shi,[6] Rasmus Nielsen,[7,*] Shuai Cheng Li,[2,11,*] and Shengbin Li[1,*]

[1]Bio-evidence Sciences Academy, Xi'an Jiaotong University, Xi'an 710061 Shaanxi, P.R. China
[2]Department of Computer Science, City University of Hong Kong, Kowloon 999077, Hong Kong
[3]Department of Biology, University of Richmond, Richmond, VA 23173, USA
[4]West China School of Basic Medicine Sciences & Forensic Medicine, Sichuan University, Chengdu 610064, Sichuan, P.R. China
[5]School of Forensic Medicine, Kunming Medical University, Kunming 650500, Yunnan, P.R. China
[6]Department of History, Sichuan University, Chengdu 610064, Sichuan, P.R. China
[7]Department of Integrative Biology, University of California, Berkeley, Berkeley, CA 94720-3140, USA
[8]Shenzhen Byoryn Technology, Shenzhen, Guangdong, P.R. China
[9]Department of Cardiology, Zhuhai People's Hospital (Zhuhai Hospital Affiliated with Jinan University), Zhuhai, Guangdong, P.R. China
[10]These authors contributed equally
[11]Lead contact
*Correspondence: rasmus_nielsen@berkeley.edu (R.N.), shuaicli@cityu.edu.hk (S.C.L.), shengbinlee@mail.xjtu.edu.cn (S.L.)
https://doi.org/10.1016/j.celrep.2022.110720

SUMMARY

The Tibetan-Yi Corridor (TYC) region between Tibet and the rest of east Asia has served as a crossroads for human migrations for thousands of years. The lack of whole-genome sequencing data specific to the TYC populations has hindered the understanding of the fundamental patterns of migration and divergence between humans in east Asia and southeast Asia. Here, we provide 248 individual whole genomes from the 16 TYC and 3 outgroup populations to elucidate historical relationships. We find that the Tibetan plateau forms an important barrier to gene flow, with a more Tibetan-like ancestry in northern populations and a southern east Asian-related ancestry in south populations. An isolated population, Achang, shows a prolonged isolation and genetic drift compared to other TYC populations. We also note that previous claims regarding the history and structure of TYC populations inferred by linguistics are incompatible with the genetic evidence.

INTRODUCTION

Along the eastern edge of the Tibetan Plateau live several ethnically and linguistically diverse human populations. This region has multiple rivers and mountain ranges, mostly running from north to south and with descending altitude. These topographical features create barriers to and channels for migration, resulting in many isolated regions where these diverse populations live today, likely playing a formative role in shaping the genetic background of these peoples. With a rich cultural heritage and the highest population diversity found in China, this region covers 0.88 million km^2 and 3 provinces of China (Gansu, Sichuan, and Yunnan). Of the 56 officially recognized ethnic groups of China, 20 can be found in this region, with a population size of more than 15 million people and more than 300 unique spoken languages throughout this region. Within this region, the most well-known ethnicities are the Han, Tibetan, and Yi, thus giving rise to the name of the region, the Tibetan-Yi Corridor (TYC).

TYC populations primarily speak languages belonging to the Tibeto-Burman language group, which falls within the Sino-Tibetan family. Studies have recently argued that this language family originated in northern China with farmers along the Yellow River. The Tibeto-Burman languages derive from populations that migrated west and south onto the Tibetan Plateau and into western southeast Asia, likely associated with an expansion of millet farmers (Sagart et al., 2019; Shi, 2018; Zhang et al., 2019). The TYC region is believed to play a major role in the dispersal of Tibeto-Burman languages. A historical study (Shi, 2018) describes an initial north-to-south dispersal, establishing the Tibeto-Burman language group across the region, but it also emphasizes historical documentation of multiple different population influxes historically from the west (Tibetans), north (Mongolians), and east (Han), as well as movements within the TYC region (i.e., expansion of the Yi northward). These linguistic and historical studies suggest that the biological roots of different TYC populations are quite complex (Shi, 2018; Van Driem, 2002).

The location of the TYC in the foothills of the Tibetan Plateau underlies the role of the region as a boundary between Tibetan and Han populations, who share a close relationship and likely originated from a shared ancestral population in east Asia (Yi et al., 2010). Genomic studies have greatly affected our

understanding of populations worldwide (Ayub et al., 2015; Bergstrom et al., 2017; Fiorito et al., 2016; Francioli et al., 2014; Leslie et al., 2015; Lipson et al., 2014; Malaspinas et al., 2016; Moltke et al., 2015; Morenoestrada et al., 2014; Patin et al., 2017; Reich et al., 2009), and they can play a decisive role in clarifying the genetic structure underlying TYC populations today. Studies of whole genomes from the TYC region have thus far largely been focused on articulating the contribution of Tibetan and Han ancestries to this region. Yao et al. (Yao et al., 2017) analyzed genome-wide single nucleotide polymorphisms (SNPs) from 10 Tibetans and 10 Han Chinese from the northern TYC region for their genome-wide SNP data. It concluded that TYC populations are a mixture of ancestry related to Tibetans on the Plateau and surrounding lowland east Asians. With the diversity of ethnic and linguistic groups found in the TYC region, genome-wide studies sampling across multiple ethnic groups within this region is needed to capture the genetic patterns that underlie the cultural diversity in TYC populations.

In this work, we collected whole blood from 248 participants from the TYC region (242 individuals) and northern China (6 individuals) for whole-genome sequencing (Figure 1A; Table 1). The individuals from northern China were sequenced to a depth of 25×. For the individuals from the TYC region, 32 of them were sequenced to a depth of 25×. The other 210 individuals were sequenced to a depth of 5×. Variants were called using standard procedures (see Method details), and the number of variants is shown in Figure 1B. We compared the TYC genomic data to ~2,000 whole-genome sequences from previously published large-scale projects, including the 1000 Genomes Project (Sudmant et al., 2015), the Simons Genome Diversity Project (Mallick et al., 2016), and additional Tibetan and Han data (Lu et al., 2016). A careful analysis of their genomes, together with available genome-wide data, will provide further insights into the genetic structure of humans in the TYC region.

RESULTS

TYC populations are genetically closest to east Asians

To understand the genetic relationship between the TYC populations and other groups, we performed principal-component analysis (PCA) on a global dataset that included the TYC populations and previously published humans (Figures 2A, S1A, and S1B). PCAs based on linkage disequilibrium (LD)-independent SNPs and haplotypes both show that the TYC populations clustered with other east Asians. Similarly, the fixation index (F_{st}) values between TYC populations and other east Asians are less than 0.15, which indicates that the degree of genetic differentiation between TYC populations and other east Asians is relatively modest (Figure S1C). The haplotype-based fineSTRUCTURE analysis also shows the TYC populations, and east Asians are clustered together (Figure S2). Finally, we further confirm that the TYC populations share the closest genetic relationship to east Asians according to the outgroup-F_3 statistic results (Figure S3A).

Ancestry analysis with ADMIXTURE suggests that present-day TYC populations share the majority of their ancestry makeup with populations from east Asia and have minor ancestral relationships with central Asia, the Americas, and western Eurasia (Figure 2B). The varying proportions of Tibetan-related (Figure 2B, dark green), Dai-related (light green), central Asian-related (rose red), western Eurasian-related (brown), and TYC-related (light yellow) indicate high genetic diversity within the TYC region, even within populations. These five components cluster on the same branch of the ancestry components tree (Figure S4A). Four Tibetan subgroups in the northern TYC (Muya, Jiarong, Namuyi, and Tibetan(SC) populations) contain mostly the Tibetan-related ancestry component. In contrast, three populations inhabiting the TYC southernmost (Lahu, Jino, and Hani populations), contain a more than 50% Dai-related ancestry component. In addition, the Qiang, Bai, and Yi populations, residing at the eastern edge of the TYC, have similar genetic ancestry compositions with Han. We next consider the genetic structure across these populations in finer detail.

Characterizing Tibetan-related ancestry in officially designated Tibetan subgroups

In the regional-scale ADMIXTURE analysis (Figure 3A), several TYC populations consistently share a component with highland Tibetans (red), particularly those officially designated as Tibetans (i.e., Tibetan(SC), Muya, Ersu, Pumi(N), Namuyi, and Jiarong populations). Meanwhile, we observed that a TYC-related component (dark green), associated with these northern TYC populations, was found in some eastern Tibetan highlanders as well. These Tibetan subgroups are referred to as the Khams (or Khamba) by Tibetans and lead different lifestyles and speak different languages from Tibetan highlanders. Their interaction with the central region of the Tibetan Plateau has been at times fractious (Neumaier-Dargyay, 1997). Here, we explore further the extent to which different Tibetan-assigned TYC populations share a genetic history with highland Tibetan populations.

Regional-scale PCA showed that TYC populations distribute along the PC1 axis (from left to right) according to their geographic location (from north to south), except for Achang (Figure 3B). Tibetan(SC), Muya, Jiarong, Namuyi, and Pumi(N) populations residing in the south of the TYC, tend to cluster with Tibetan highlanders (Figures S4B–S4G). The outgroup-F_3 test demonstrated that these five Tibetan subgroups shared the closest genetic relationship to Tibetan highlanders (Figure S3B). We also used the outgroup-F_3 statistic and the D-statistic to determine whether each TYC population shared higher genetic similarity to Tibetans or the southern east Asian Dai. We observed that the Muya, Jiarong, and Tibetan(SC) populations showed much higher genetic similarity to Tibetans than to Dai, and the Namuyi, Pumi(N), Pumi, and Ersu populations also showed similar patterns, although less extreme (Figure 3C; Tables S1A, S1B, and S1C). However, Ersu and Pumi(N) populations harbor a closer genetic relationship with Han than Tibetans (Figure 3C) and share a similar number of alleles to both southern TYC populations and Tibetans (Table S1D). Using ChromoPainter and fineSTRUCTURE, we constructed dendrograms that clustered populations from the TYC and nearby regions. Muya, Jiarong, Tibetan(SC), Namuyi, and one each of the Pumi and Pumi(N) clustered with Tibetans, while the Ersu did not (Figure 3D). These results suggest that ancestry not related to Tibetans has played a role in the genetic history of the

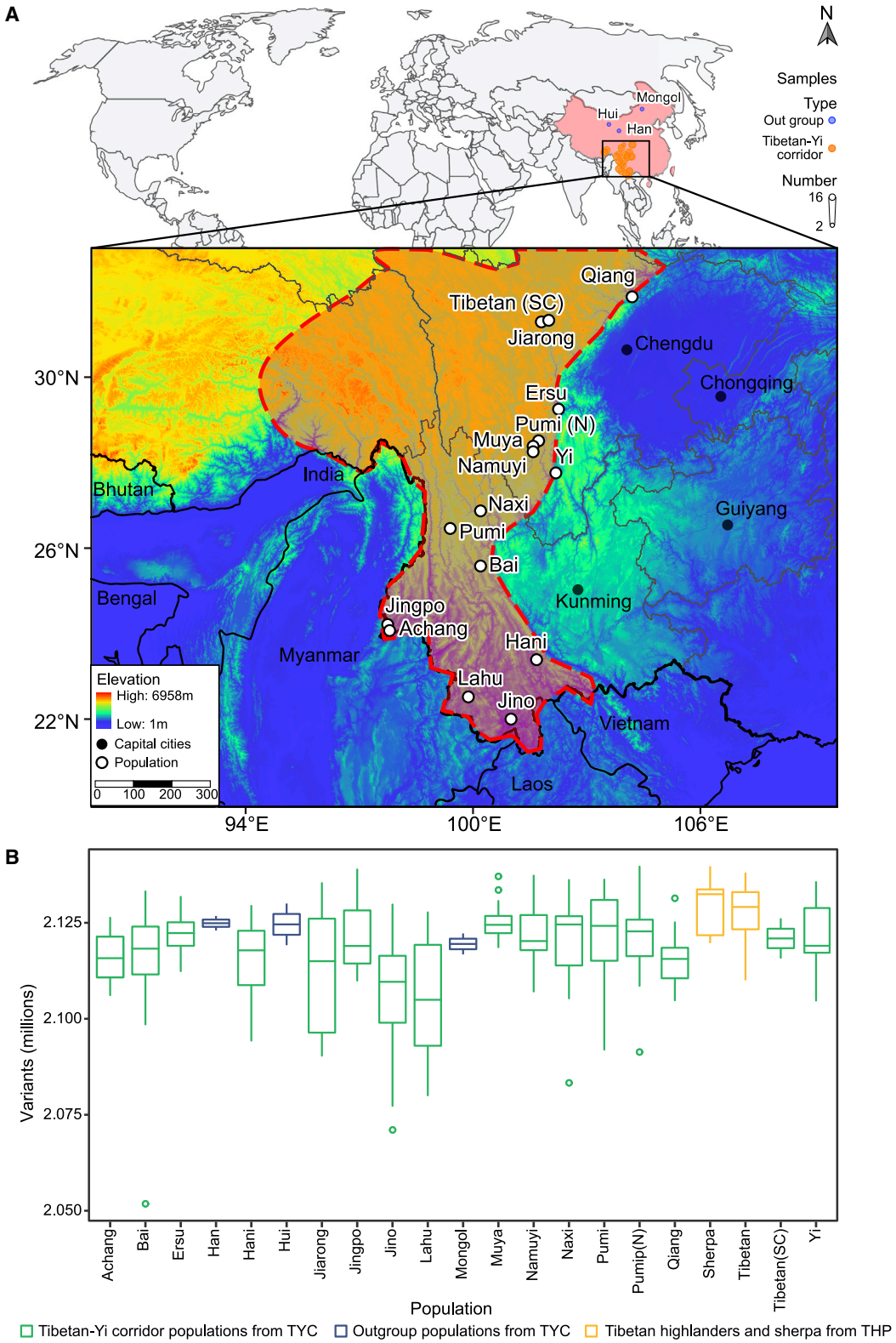


Figure 1. TYC population distribution and variants analysis

(A) Map of sampled populations in the Tibetan-Yi Project (detailed in Table 1). The area within the red dotted line is the Tibetan-Yi Corridor (TYC).

(B) A boxplot showing the distribution of variants for different populations in TYC.

Table 1. Sample information

Population		No. sample	Sequencing depth	Language[a]	Altitude (m)	Latitude	Longitude
Tibetan-Yi Corridor population	Tibetan(SC)	2	25×	WTB, Central Bodish, Tibetan, Khams	2,137	31.11	101.88
	Jiarong	2	25×	NTB, rGyalrongic, Jiarong	2,066	31.12	101.89
		14	5×				
	Ersu	2	25×	NTB, Ersuish	1,730	29.20	102.25
		14	5×				
	Pumi	2	25×	NTB, Qiangic, Pumi Northern	2,631	29.62	101.61
		14	5×				
	Muya	2	25×	NTB, Qiangic, Muya	3,210	29.62	101.61
		14	5×				
	Namuyi	2	25×	NTB, Naic, Namuyi	2,631	28.31	101.61
		14	5×				
	Yi	2	25×	NB, Loloish, Yi	2,114	27.74	102.27
		14	5×				
	Qiang	2	25×	NTB, Qiangic, Qiang	1,193	31.82	104.18
		14	5×				
	Naxi	2	25×	NTB, Naic, Naxi	2,651	26.82	100.24
		14	5×				
	Pumi	2	25×	NTB, Qiangic, Pumi, northern	3,065	26.45	99.42
		14	5×				
	Bai	2	25×	NTB, Bai	1,993	25.58	100.23
		14	5×				
	Lahu	2	25×	NB, Loloish, Lahu	1,363	22.57	99.93
		14	5×				
	Jinuo	2	25×	NB, Loloish, Jinuo	945	22.04	101.01
		14	5×				
	Jingpo	2	25×	Sal, Jingphaw, Jingpho	1,155	24.18	97.79
		14	5×				
	Achang	2	25×	NB, Burmish, Achang	962	24.03	97.79
		14	5×				
	Hani	2	25×	NB, Loloish, Hani	1,145	23.43	101.69
		14	5×				
Outgroup	Han	2	25×	Chinese	392	34.34	108.94
	Hui	2	25×	Chinese	2,132	35.60	103.24
	Mongol	2	25×	Altaic, Mongolic, Mongolian	181	43.62	122.26
Total		24					

[a]The languages are classified according to the Ethnologue database (http://www.ethnologue.com). NB, Ngwi-Burmese; NTB, northeastern Tibeto-Burman; WTB, western Tibeto-Burman.

Pumi(N), Pumi, and Ersu, but that they also have a close relationship to the Tibetan-like populations in the TYC region—Muya, Jiarong, and Tibetan(SC). These results showed that the northern TYC populations shared the ancestry component with highland Tibetans, but local genetic drift also occurred within the northern TYC.

Southern east Asian ancestry in the southern TYC populations

The close genetic relationships between southern TYC populations (i.e., Lahu, Jino, and Hani) and southern east Asians (e.g., Dai, Thai, Cambodians, and Kinh) in our ADMIXTURE analysis (Figure 2B) highlighted that an ancient population unrelated to the migration of the Tibeto-Burman spoken language but perhaps related to an ancient southern east Asian ancestry (Yang et al., 2020) played a major role in the peopling of the region of south TYC. In the outgroup-F_3 statistic, Lahu, Jino, and Hani populations show high genetic similarity with southern east Asians (Figures 3C and S3B). Cambodians are closer in ancestry to these southern TYC populations than all of the other TYC populations (Table S1E). According to the literature, the expansion of the Nanzhao Kingdom southward is thought to

CellPress
OPEN ACCESS

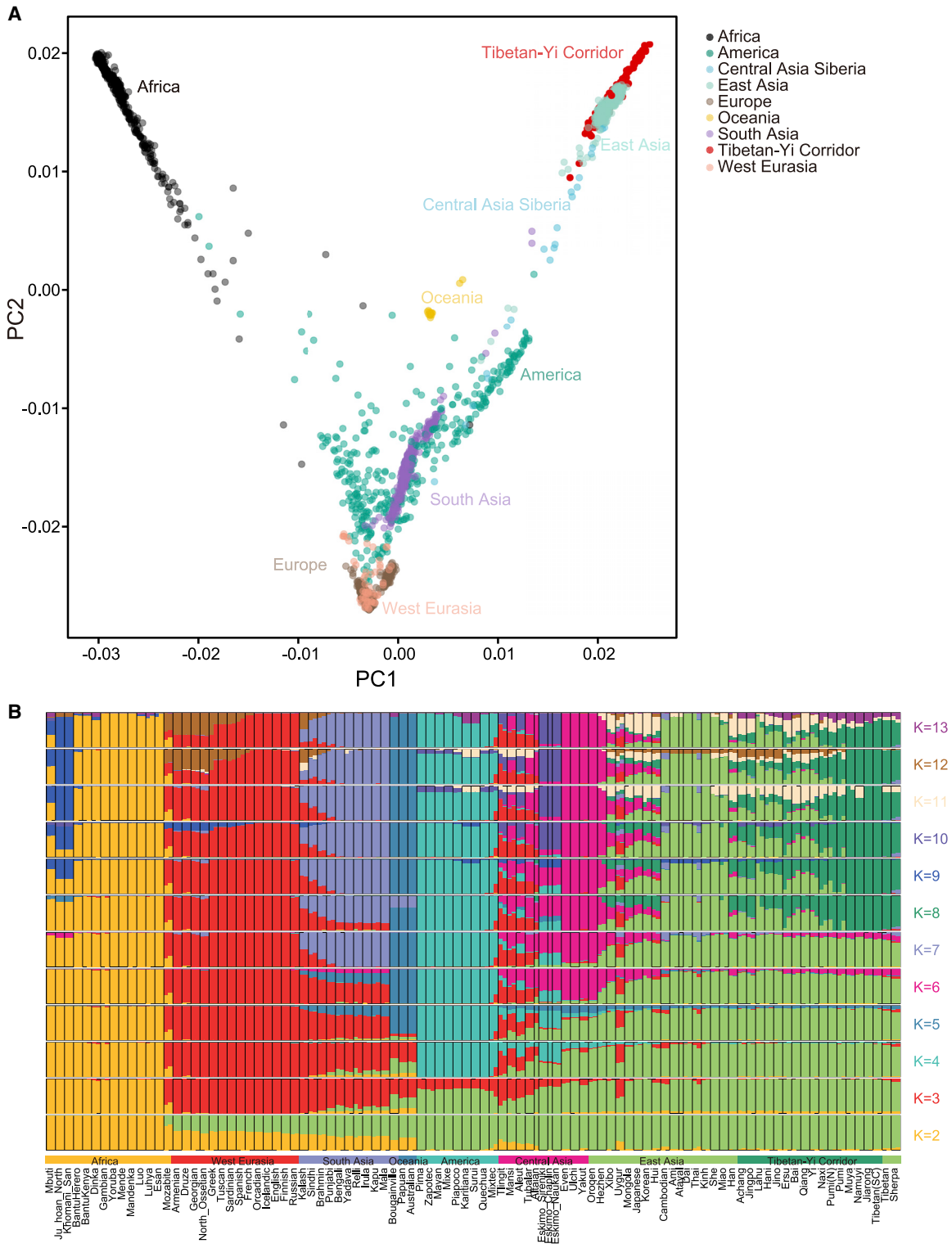


Figure 2. Population stratification and genetic structure analysis

(A) PCA of a wide range of populations (TYC, SGDP, 1KG, and THP). 1KG, 1000 Genomes Project Phase 3; SGDP, Simons Genome Diversity Project; THP, Tibetan Highlanders Project.

(B) ADMIXTURE results for high-coverage individuals, with k = 2–13.

See also Figures S1–S4.

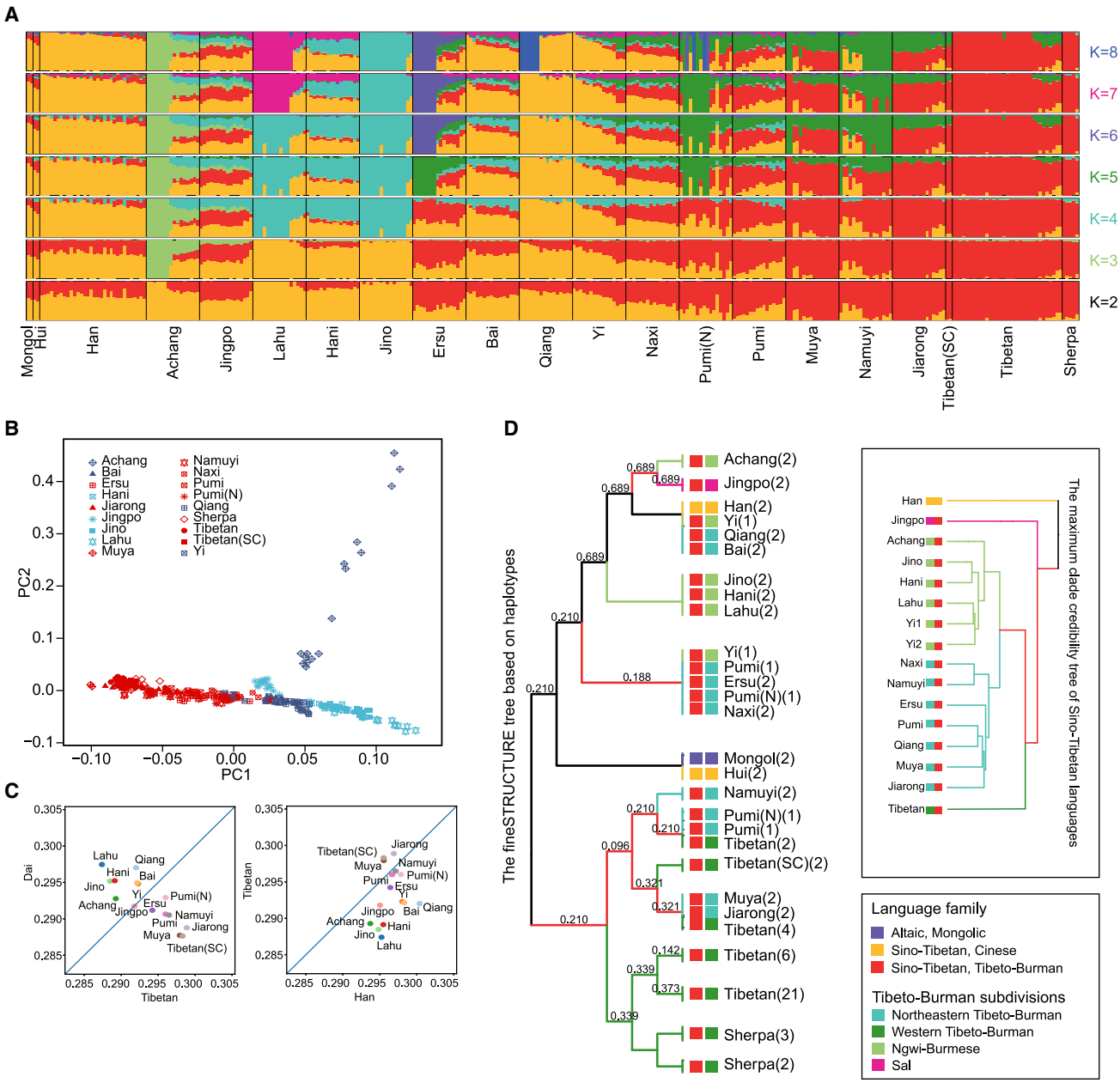


Figure 3. Genetic differences within TYC populations

(A) ADMIXTURE results for all individuals, with k = 2–6.

(B) PCA of TYC populations.

(C) The scatter diagram for the outgroup-F_3 test in the form of f_3(Mbuti; TYC, Tibetan), f_3(Mbuti; TYC, Han), and f_3(Mbuti; TYC, Dai). Left: Outgroup-F_3 test in the form of f_3(Mbuti; X, Dai) and f_3(Mbuti; X, Tibetan), where X represents a population on TYC. Dots above the diagonal line mean that the populations have a closer relationship to Dai compared to Tibetan, dots below the diagonal line mean that the populations have a closer relationship to Tibetan compared to Dai, and dots that stand on the line represent the populations that are the same, close to Tibetan and Dai. Right: Outgroup-F_3 test in the form of f_3(Mbuti; X, Tibetan) and f_3(Mbuti; X, Han), where X represents a population on TYC. Dots above the diagonal line mean that the populations have a closer relationship to Tibetan compared to Han, dots below the diagonal line mean that the populations have a closer relationship to Han compared to Tibetan, and dots that stand on the line represent the populations that are the same, close to Han and Tibetan.

(D) Clustering trees based on co-ancestry matrices of counts of shared haplotypes. The dendrogram shows the inferred relationship between populations. Linguistic groups are indicated with colors as shown in the legend. The numbers in the brackets indicate the individual numbers of each ethnicity group. The languages are classified according to the Ethnologue database (http://www.ethnologue.com). The maximum clade credibility tree of Sino-Tibetan languages was adapted from Zhang et al. (2019).

See also Figures S4B–S4G.

have brought Tibeto-Burman languages into Southeast Asia (Lapolla, 2013). Our study highlights a common shared ancestry in TYC southernmost, other regions of southern China, and southeast Asia that differs from that in northern TYC and the Tibetan Plateau. The widespread genetic commonalities may be tied in part to the expansion of the Nanzhao Kingdom.

Han-like ancestry compositions in the eastern TYC edge populations

One of the dominant ethnic groups in the TYC region is the Han, largely from increased migration to this area in recent history. Starting in the Ming Dynasty (14th–17th centuries), the TYC region began to be a bridge between the Tibetan Plateau and the more eastern populations on the plains of China. During the Qing Dynasty, from the 18th to 20th centuries, more Han people migrated to this region due to increased interactions and increased defenses on the western border of China (Shi, 2018). ADMIXTURE analyses above have shown that the ancestry components of Qiang, Yi, Bai, and Han populations resembled each other (Figure 3A). The fineSTRUCTURE analysis showed that the Qiang, Yi, Bai, and Han populations were clustered on the same branch of the clustering tree. PCA, outgroup-F_3 statistic, and D-statistic results also showed that these three populations, unlike the other TYC populations, share the closest genetic relationship to Han (Figures 3C, S3B, and S4E–S4G).

The Yi population is one of the largest ethnic groups in the TYC region, with a rich and ancient history tied to the region (Gu, 2001). They expanded northward in the last 2 centuries, greatly increasing the movement and location of Yi populations today. Previous genetic studies have shown what we observed, that the Yi have both Tibetan and Han connections (Gu, 2001; Shi, 2018; Wang et al., 2011). The Bai, who also has a rich history local to the TYC region, has a history of acculturation with the Han (Mackerras, 1988). Here, we found that cultural assimilation of the Bai and northern expansion of the Yi are both associated with increasing Hans. The Qiang have been described in the Chinese literature as early as the Shang Dynasty as people to the west of the Han. Later references in the 5th century describe them as inhabitants of the Hehuang region, which spans the upper Yellow River valley and Huang River valley (Wang, 2002). Mitochondrial haplotypes showed that the Qiang had the greatest diversity of Sino-Tibetan populations, supporting data from historical and archaeological studies indicating that the Qiang group is the origin of the Sino-Tibetan expansion. Our findings do not exclude this hypothesis.

High genetic drift in the Achang populations

The Achang population is considered to be early inhabitants of Yunnan (Jiakai, 2003). In the ADMIXTURE analysis, the Achang consistently kept a unique ancestry component, also found in the Jingpo (Figure 3A). The Achang population is scattered along the PC2 axis and isolated from other populations, except for the Jingpo population (Figure 3B). The fineSTRUCTURE analysis shows that the Achang and Jingpo populations are clustered under a separate branch (Figure 3D). The Achang population also has a distant genetic relationship with other TYC populations, except for Jingpo, according to the outgroup-F_3 statistic results (Figure S3B). In the D-statistic symmetry test, the Achang population consistently shared more affinity to the Jingpo population (Table S1F). The closest genetic relationship between Achang and Jingpo is likely due to the neighboring geographic locations. Meanwhile, these results suggest that the Achang experienced high genetic drift, differentiating them from other TYC populations.

We compared runs of homozygosity (ROH), LD, and heterozygosity in the TYC populations to further explore this issue. The Achang do not have particularly long ROHs (Figure 4A), suggesting that they are not highly inbred. However, the Achang has the most significant LD (Figure 4B) and the least heterozygous locus (Figure 4C). These results likely partially explain the high divergence of the Achang population and are consistent with their relatively small census population size (40,000) of the China Census in 2010 (http://www.stats.gov.cn/tjsj/pcsj/rkpc/6rp/indexch.htm). In conjunction with the multiple sequentially Markovian coalescent (MSMC) results (Figure S5), we suggest that the Achang population lives in relative isolation for a prolonged period with relatively small N_e, resulting in the increased genetic drift.

Inconsistent between linguistics and genetic affinities

Haplotype-based clustering trees of TYC populations (Figure 3D) show that populations from the same linguistic group tend to cluster together. Nevertheless, we identified eight population outliers with inconsistent linguistic and genetic affinities (i.e., Jingpo, Achang, and Yi), and five Tibetan subgroups. Specifically, Tibetan subgroups (i.e., Muya, Ersu, Pumi(N), Namuyi, and Jiarong populations) shared the ancestry component with highland Tibetans but spoke different languages. Jingpho is an independent Sal branch of the Tibeto-Burman language spoken by Jingpo, who has a close genetic relationship with Achang. Jino, Hani, Lahu, and Yi populations, were located in different branches, but their language belongs to the northeastern Tibeto-Burman.

Barriers to gene flow

To further understand historical patterns of gene flow in the region, we constructed a map to visualize the ancestry components in a geographical context (Figures 5A and S6). We also estimated migration barriers using estimating effective migration surfaces (EEMS) and isolation by distance with an isotropic migration model (Figure 5B) (Petkova et al., 2016). These results illustrate that the slopes of the Tibetan plateau likely have been an important barrier to gene flow between eastern east Asia and the Tibetan Plateau. Using a stricter threshold in the EEMS analyses of 0.9, another barrier to gene flow emerges in the southeastern Yunnan region (Figure 5C). The migration barriers and pathways to gene flow are in a north-to-south direction. Moreover, admixture-F_3 statistic results show that the populations living in the southern part of the TYC are derived from an admixture of populations from northern TYC, suggesting that these admixed populations resulted from north-to-south migration events (Table S2).

Historical, linguistic, and archaeological studies have suggested a north-to-south movement of populations from the upper reaches of the Yellow River to the TYC region, likely a population of ancestral Tibeto-Burman speakers that slowly moved southward throughout the Neolithic into southwest Yunnan and then into southeast Asia (Shi, 2018). Others (Shi, 2018; Shuo, 2008) have also found evidence suggesting migration from north

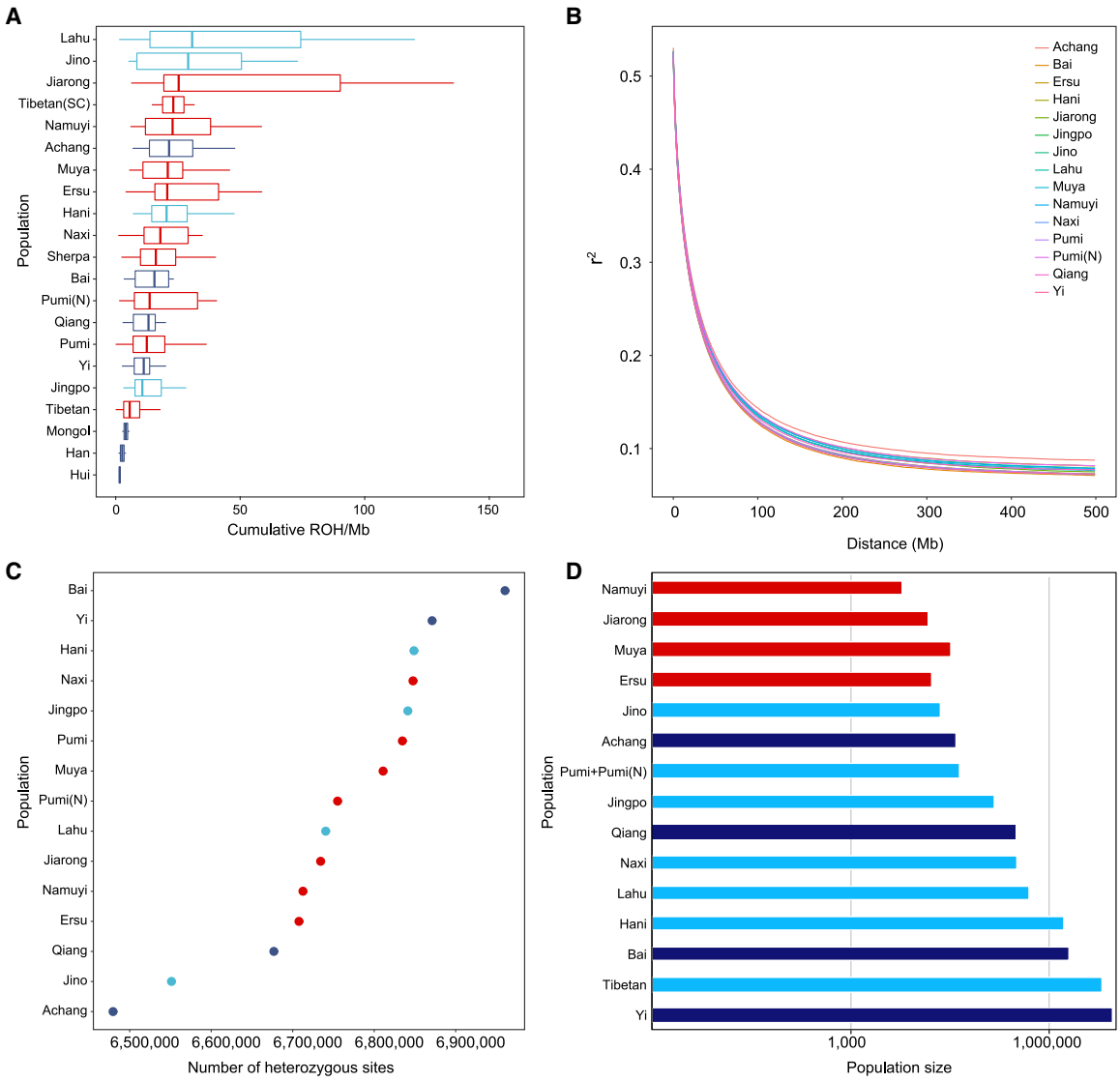


Figure 4. Genetic studies for TYC populations

(A) Runs of homozygosity for TYC populations.

(B) Decay of linkage disequilibrium for TYC populations.

(C) Numbers of heterozygous sites in TYC populations.

(D) Population sizes of TYC populations.

See also Figure S5.

to south in the TYC region. Our findings do not exclude this hypothesis, but they add complexity to a simple north-to-south narrative. Together, these results highlight the importance of the TYC region, not only as a cultural transition zone between the Tibetan Plateau and more eastern populations but also as a biological transition zone.

DISCUSSION

The TYC region is a meeting point between many east Asian populations, especially those closely related to the Han, Tibetan, and southern east Asians. Our genetic results highlight the complexity of their population history, with a similarly complex admixture. However, some clear patterns can be observed. First, one group (Tibetan(SC), Jiarong, Muya) consistently share a close relationship and a close relationship with highland Tibetans. The Pumi, Namuyi, Pumi(N), Naxi, and Ersu, who live along the fringes of the Tibetan Plateau, share a close relationship with this group. However, their affinity to Tibetans is weaker, and they share genetic drift, representing a northern TYC-specific ancestry. Second, the Qiang, Bai, and Yi show high connections to the Han, indicating that they are heavily admixed and making it difficult to assess their ancestral origins before admixture with Han-related populations. Third, the Achang show high

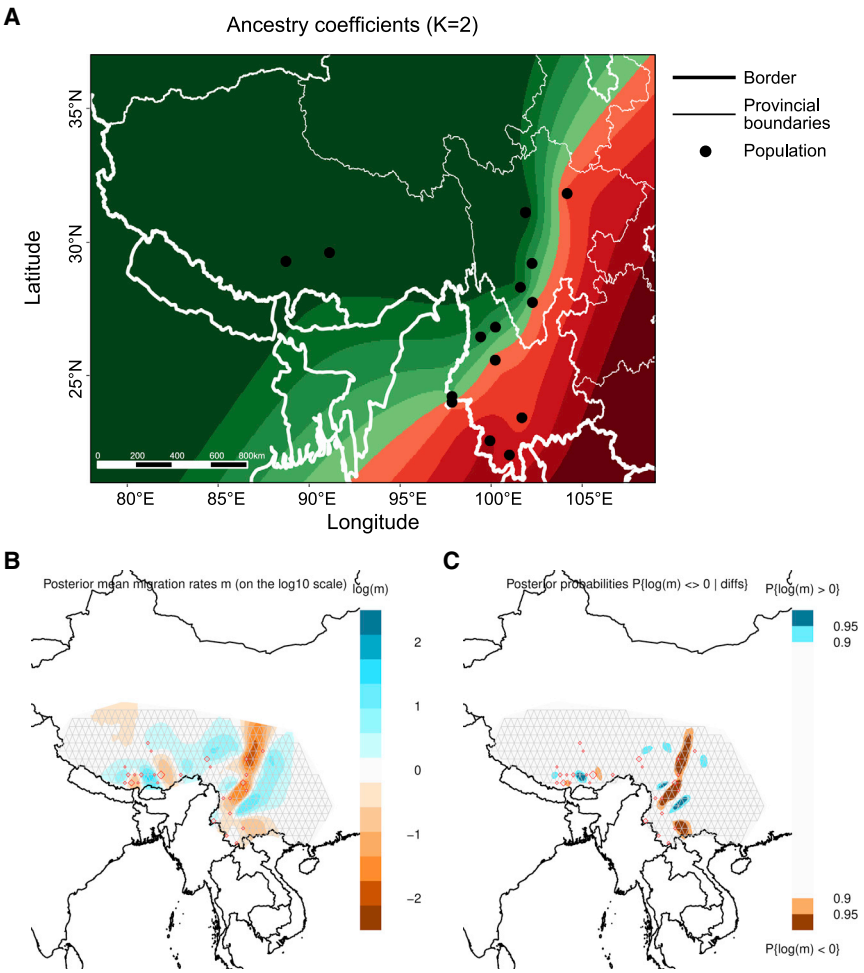


Figure 5. Migration barriers and channels analysis

(A) Interpolated ancestry coefficients on the geographic map for k = 2. A clear cline appears from east to west, likely representing Tibetan-like ancestry.

(B) Effective migration surfaces inferred by EEMS, with the threshold as 0.1. We identify a clear barrier (log(m) < 0) to gene flow between east and west, where the barrier falls along the TYC region.

(C) Effective migration surfaces inferred by EEMS, with the threshold as 0.9.

See also Figure S6.

genetic drift, differentiating them from other TYC populations, likely due to their small population size. They share an affinity to southern east Asians and the neighboring Jingpo. Finally, the Hani, Lahu, and Jino in the south of the TYC region are predominantly of southern east Asian ancestry.

Culturally, there are shared customs across the northern and southern TYC regions. Many TYC populations practice a funeral rite called "escorting the soul" (Shi, 2018). During the funeral ceremony, priests "send" the soul to their ancestral land by listing previous locations their ancestors lived. While many location names are unidentifiable today, researchers have found that all of the routes led to the north, suggesting a northern origin of these populations (Stevan, 2001). Seven TYC populations have this custom (Yi, Naxi, Hani, Lahu, Jino, Jingpo, and Pumi), and aspects of this rite are also partially executed in the Achang and Bai (Yuan and Chen, 2011). Notably, these do not include populations with high Tibetan-related ancestry, but they do include both northern and southern TYC populations, illustrating shared customs between north and south TYC populations. We observe distinct ancestry clusters in the north and south TYC regions that do not show a northern origin for southern TYC populations (except perhaps Jingpo). Still, we find high levels of admixture, indicating interaction across the TYC region.

The distinct grouping of southern populations suggests that southern east Asian ancestries may have also contributed to populations in the TYC corridor. Its widespread presence in this region and southeast Asia may be tied to the expansion of the Nanzhao Kingdom in the 8th century AD, but ultimately, whether it originated from a prehistoric north-to-south migration from the upper reaches of the Yellow River is difficult to discern due to the high amount of admixture from more recent historical events. Compared to the overlapping linguistic phylogeny adapted from Zhang et al. (2019), we similarly observe that most of the northern and southern TYC populations cluster along geographic lines (Figure 3D). However, we note many differences between the linguistic and genetic relationships, highlighting dissonant cultural and genetic patterns. The Qiang and Yi show a close genetic relationship to the Han, but linguistically, they group separately into the northeastern Tibeto-Burman and Ngwi-Burman clusters, respectively. Although Tibetans, Jiarong, and Muya are genetically closely related, the Tibetan language is an outgroup to the northern and southern TYC language groups. Unusually, the Jingpo, who share a genetic relationship with Achang, belongs to a language outgroup that is highly diverged. This linguistic outlier clusters with their geographic neighbors. The same phenomenon was earlier found in the study of the HUGO Pan-Asian SNP Consortium (Abdulla et al., 2009). These inconsistencies between linguistics and genetic affinities patterns could be due to either substantial recent admixture among the populations, a history of language replacement, or uncertainties in the linguistic classifications themselves. These conflicts between the linguistic and genetic relationships highlight that a closer examination of the genetic and cultural patterns in TYC populations is needed.

The TYC region is the key to understanding the fundamental migration patterns and divergence between humans in east Asia and southeast Asia. Our genetic results demonstrated the complexity of their population history with some clear patterns.

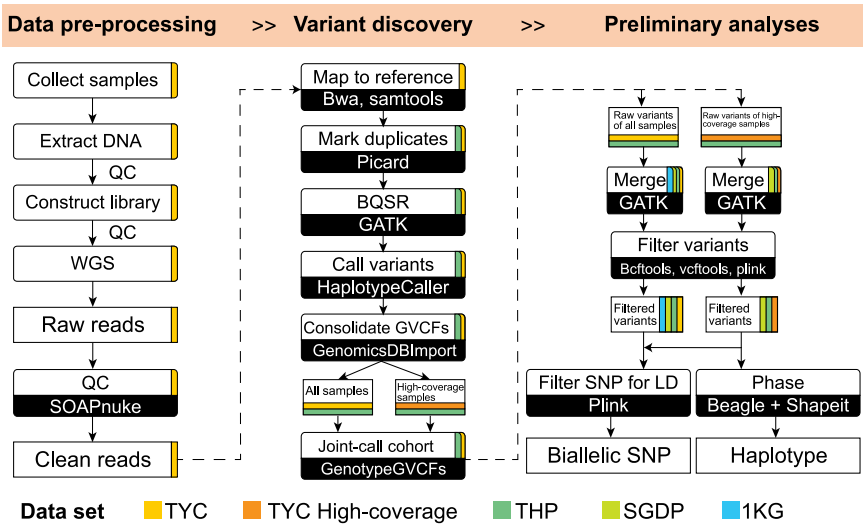


Figure 6. Summary of the callset generation pipeline

Rectangles, input or output data, round-corner rectangles, analytical methods, and software. The small colored box on the right side of each box corresponds to the dataset used in this step. GATK, Genome Analysis ToolKit; GVCF, Genomic Variant Call Format; QC, quality control; WGS, whole-genome sequencing; BQSR, base quality score recalibration.

A new genetic ancestry related to but distinct from the current Tibetan ancestry is discovered in this study. We show that there are understudied and very distinct population isolates in this region, such as the Achang people. We also show that the general pattern of genetic variation in the region mirrors the geography, with the slopes of the Himalayas as a major barrier to gene flow. The genetic analysis portrays a gradual change in genetic relationships from north to south. However, it does not always agree with the phylogenetic analysis of people in the region solely based on linguistic evidence. Rapid development today is leading to a swift loss of languages and cultures from this region. More study of this region is needed to rescue the languages, culture, heritage, and biology of people of the TYC.

Limitations of the study

Genetic patterns do not always agree with evidence from language or history, likely because populations heavily interacted, leading to highly admixed populations. This study performed deep sequencing on 2–3 samples and shallow-depth sequencing on 14 samples for each TYC population (Figure 6). Sequencing more minority samples with high coverage is still useful to explore the complex genetic patterns in TYC populations.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
- METHOD DETAILS
  - DNA extraction, library construction, and sequencing
  - Alignment and variant calling
  - TYC sequencing data and previously published data combining
  - Site quality filtering
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Kinship analysis
  - Haplotype-based analysis
  - Principal component analysis (PCA)
  - Model-based clustering
  - Genetic differences between populations (F_{st})
  - F_3 statistics and D-statistics analyses
  - Runs of homozygosity
  - SFS estimation
  - Historical population effective size inference
  - Migration and isolation by distance

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.celrep.2022.110720.

ACKNOWLEDGMENTS

The results published here are wholly or in part based upon data generated by the Tibetan-Yi Corridor Project. It was fully supported by a grant of Xi'an Jiaotong University, P.R. China, to Shengbin Li. (no. 2013FY114300). We would also like to thank Dr. Jun Yu for valuable comments and critical reading of the manuscript. Finally, we wish to thank the participants and their families for their contributions to valuable data, without which this project would not have been possible.

AUTHOR CONTRIBUTIONS

Shengbin Li coordinated the study. S.L., Shuai Cheng Li, and Z.Z. conceived the study. Z.Z. coordinated and Y.W., B.Z., L.Z., B.X., H.Z., and T.C. collected the samples. Zhe Zhang and Y.W. supervised data acquisition and delivery. R.N. advised on the data analyses and interpretation. Shuai Cheng Li designed and performed most of the data analyses. Yanlin Zhang, W.D., and Yong Zhou performed the data processing and wrote the computer program. Zicheng Zhao performed the mutation callings and the F_3 statistics. S.S. provided

deep insights into the data interpretation from the perspective of cultures, heritage, and anthropology. Zhe Zhang, Yanlin Zhang, Y.W., Zicheng Zhao, R.N., Shengbin Li, M.Y., and Shuai Cheng Li wrote the manuscript and formatted all of the figures. All of the authors have proofread the manuscripts.

DECLARATION OF INTERESTS

The authors declare no competing interests.

INCLUSION AND DIVERSITY

We worked to ensure gender balance in the recruitment of human subjects. We worked to ensure that the study questionnaires were prepared in an inclusive way. The author list of this paper includes contributors from the location where the research was conducted who participated in the data collection, design, analysis, and/or interpretation of the work.

Received: March 1, 2021
Revised: November 8, 2021
Accepted: March 31, 2022
Published: April 26, 2022

REFERENCES

1000 Genomes Project Consortium, Auton, A., Brooks, L.D., Durbin, R.M., Garrison, E.P., Kang, H.M., Korbel, J.O., Marchini, J.L., McCarthy, S., McVean, G.A., and Abecasis, G.R. (2015). A global reference for human genetic variation. Nature *526*, 68.

Abdulla, M.A., Ahmed, I., Assawamakin, A., Bhak, J., Brahmachari, S.K., Calacal, G.C., Chaurasia, A., Chen, C.H., Chen, J., Chen, Y.T., et al. (2009). Mapping human genetic diversity in Asia. Science *326*, 1541–1545.

Alexander, D., Novembre, J., and Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. Genome Res. *19*, 1655–1664.

Auton, A., Brooks, L.D., Durbin, R.M., Garrison, E.P., Kang, H.M., Korbel, J.O., Marchini, J.L., McCarthy, S., McVean, G.A., and Abecasis, G.R. (2015). A global reference for human genetic variation. Nature *526*, 68–74.

Ayub, Q., Mezzavilla, M., Pagani, L., Haber, M., Mohyuddin, A., Khaliq, S., Mehdi, S.Q., and Tylersmith, C. (2015). The Kalash genetic isolate: ancient divergence, drift, and selection. Am. J. Hum. Genet. *96*, 775–783.

Baroud, G., and Steffen, T. (2005). A new cannula to ease cement injection during vertebroplasty. Eur. Spine J. *14*, 474–479.

Bergstrom, A., Oppenheimer, S.J., Mentzer, A.J., Auckland, K., Robson, K.J.H., Attenborough, R., Alpers, M.P., Koki, G., Pomat, W., and Siba, P. (2017). A Neolithic expansion, but strong genetic structure, in the independent history of New Guinea. Science *357*, 1160–1163.

Broad Institute (2019). Picard Toolkit (GitHub Repository). https://broadinstitute.github.io/picard/.

Browning, S.R., and Browning, B.L. (2007). Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. Am. J. Hum. Genet. *81*, 1084–1097.

Chang, C.C., Chow, C.C., Tellier, L.C., Vattikuti, S., Purcell, S.M., and Lee, J.J. (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. GigaScience *4*, 7.

Chen, Y., Chen, Y., Shi, C., Huang, Z., Zhang, Y., Li, S., Li, Y., Ye, J., Yu, C., Li, Z., et al. (2018). SOAPnuke: a MapReduce acceleration-supported software for integrated quality control and preprocessing of high-throughput sequencing data. GigaScience *7*, 1–6.

Cheng, J.Y., Mailund, T., and Nielsen, R. (2016). Ohana, a tool set for population genetic analyses of admixture components. Preprint at bioRxiv. https://doi.org/10.1101/071233.

Cheng, J.Y., Mailund, T., and Nielsen, R. (2017). Fast admixture analysis and population tree estimation for SNP and NGS data. Bioinformatics (Oxford, England) *33*, 2148–2155.

Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., DePristo, M.A., et al. (2011). The variant call format and VCFtools. Bioinformatics *27*, 2156–2158.

Delaneau, O., Marchini, J., and Zagury, J.F. (2012). A linear complexity phasing method for thousands of genomes. Nat. Methods *9*, 179–181.

Delaneau, O., Zagury, J.F., and Marchini, J. (2013). Improved whole-chromosome phasing for disease and population genetic studies. Nat. Methods *10*, 5–6.

DePristo, M.A., Banks, E., Poplin, R., Garimella, K.V., Maguire, J.R., Hartl, C., Philippakis, A.A., del Angel, G., Rivas, M.A., Hanna, M., et al. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat. Genet. *43*, 491–498.

Van Driem, G. (2002). Tibeto-Burman replaces Indo-Chinese in the 1990s: review of a decade of scholarship. Lingua *112*, 79–102.

Fiorito, G., Gaetano, C.D., Guarrera, S., Rosa, F., Feldman, M.W., Piazza, A., and Matullo, G. (2016). The Italian genome reflects the history of Europe and the Mediterranean basin. Eur. J. Hum. Genet. *24*, 1056–1062.

Francioli, L.C., Menelaou, A., Pulit, S.L., Van Dijk, F., Palamara, P.F., Elbers, C.C., Neerincx, P.B.T., Ye, K., Guryev, V., and Kloosterman, W.P. (2014). Whole-genome sequence variation, population structure and demographic history of the Dutch population. Nat. Genet. *46*, 818–825.

Gu, W. (2001). Reconstructing Yi History from Yi Records (Berkeley: University of California Press), pp. 21–34.

Jiakai, D. (2003). A study in history of Achang nationality "cross the border" people [J]. J. Baoshan Teachers' Coll. *4*, 53–59.

Kopelman, N.M., Mayzel, J., Jakobsson, M., Rosenberg, N.A., and Mayrose, I. (2015). Clumpak: a program for identifying clustering modes and packaging population structure inferences across K. Mol. Ecol. Resour. *15*, 1179–1191.

Korneliussen, T.S., Albrechtsen, A., and Nielsen, R. (2014). ANGSD: analysis of next generation sequencing data. BMC Bioinformatics *15*, 356.

Lapolla, R.J. (2013). Eastern Asia: Sino-Tibetan Linguistic History (Blackwell Publishing Ltd)).

Lawson, D.J., Hellenthal, G., Myers, S., and Falush, D. (2012). Inference of population structure using dense haplotype data. PLoS Genet. *8*, e1002453.

Leslie, S., Winney, B., Hellenthal, G., Davison, D., Boumertit, A., Day, T., Hutnik, K., Royrvik, E.C., Cunliffe, B., and Lawson, D.J. (2015). The fine-scale genetic structure of the British population. Nature *519*, 309–314.

Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics *27*, 2987–2993.

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics (Oxford, England) *25*, 1754–1760.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. (2009). The sequence alignment/map format and SAMtools. Bioinformatics *25*, 2078–2079.

Lipson, M., Loh, P., Patterson, N., Moorjani, P., Ko, Y., Stoneking, M., Berger, B., and Reich, D. (2014). Reconstructing Austronesian population history in island Southeast Asia. Nat. Commun. *5*, 4689.

Lu, D., Lou, H., Yuan, K., Wang, X., Wang, Y., Zhang, C., Lu, Y., Yang, X., Deng, L., and Zhou, Y. (2016). Ancestral origins and genetic history of Tibetan highlanders. Am. J. Hum. Genet. *99*, 580–594.

Mackerras, C. (1988). Aspects of Bai culture change and continuity in a yunnan nationality. Mod. China *14*, 51–84.

Malaspinas, A., Westaway, M.C., Muller, C., Sousa, V., Lao, O., Alves, I., Bergstrom, A., Athanasiadis, G., Cheng, J.Y., and Crawford, J.E. (2016). A genomic history of Aboriginal Australia. Nature *538*, 207–214.

Mallick, S., Li, H., Lipson, M., Mathieson, I., Gymrek, M., Racimo, F., Zhao, M., Chennagiri, N., Nordenfelt, S., and Tandon, A. (2016). The Simons genome diversity project: 300 genomes from 142 diverse populations. Nature *538*, 201–206.

Manichaikul, A., Mychaleckyj, J.C., Rich, S.S., Daly, K., Sale, M., and Chen, W.M. (2010). Robust relationship inference in genome-wide association studies. Bioinformatics *26*, 2867–2873.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., et al. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. *20*, 1297–1303.

Moltke, I., Fumagalli, M., Korneliussen, T.S., Crawford, J.E., Bjerregaard, P., Jorgensen, M.E., Grarup, N., Gullov, H.C., Linneberg, A., and Pedersen, O. (2015). Uncovering the genetic history of the present-day Greenlandic population. Am. J. Hum. Genet. *96*, 54–69.

Morenoestrada, A., Gignoux, C.R., Fernandezlopez, J.C., Zakharia, F., Sikora, M., Contreras, A.V., Acunaalonzo, V., Sandoval, K., Eng, C., and Romerohidalgo, S. (2014). The genetics of Mexico recapitulates Native American substructure and affects biomedical traits. Science *344*, 1280–1285.

Neumaier-Dargyay, E. (1997). A traditional culture in transition: observations on the Tibetans in sichuan and Gansu. East Asian Cult. Hist. Perspect. Toronto, 301–312.

Patin, E., Lopez, M., Grollemund, R., Verdu, P., Harmant, C., Quach, H., Laval, G., Perry, G.H., Barreiro, L.B., and Froment, A. (2017). Dispersals and genetic adaptation of Bantu-speaking populations in Africa and North America. Science *356*, 543–546.

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

Petkova, D., Novembre, J., and Stephens, M. (2016). Visualizing spatial population structure with estimated effective migration surfaces. Nat. Genet. *48*, 94–100.

Price, A.L., Patterson, N.J., Plenge, R.M., Weinblatt, M.E., Shadick, N.A., and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. Nat. Genet. *38*, 904–909.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A., Bender, D., Maller, J., Sklar, P., de Bakker, P.I.W., Daly, M.J., and Sham, P.C. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. *81*, 559–575.

Reich, D., Thangaraj, K., Patterson, N., Price, A.L., and Singh, L. (2009). Reconstructing Indian population history. Nature *461*, 489–494.

Sagart, L., Jacques, G., Lai, Y., Ryder, R.J., Thouzeau, V., Greenhill, S.J., and List, J. (2019). Dated language phylogenies shed light on the ancestry of Sino-Tibetan. Proc. Natl. Acad. Sci. U S A. *116*, 10317–10322.

Schiffels, S., and Durbin, R. (2014). Inferring human population size and separation history from multiple genome sequences. Nat. Genet. *46*, 919–925.

Shi, S. (2018). Ethnic flows in the Tibetan-Yi corridor throughout history. Int. J. Anthropol. Ethnol. https://doi.org/10.1186/s41257-018-0009-z.

Shuo, S. (2008). The migration of people in the upper Yellow River area to the Tibetan-Yi corridor from the perspective of neolithic culture (in Chinese). J. Southwest Univ. Nationalities (Humanities Soc. Sci. Edition), 7–13.

Stevan, H. (2001). Ways of Being Ethnic in Southwest China (University of Washington Press), [Google Scholar]).

Sudmant, P.H., Rausch, T., Gardner, E.J., Handsaker, R.E., Abyzov, A., Huddleston, J., Zhang, Y., Ye, K., Jun, G., and Fritz, M.H.Y. (2015). An integrated map of structural variation in 2,504 human genomes. Nature *526*, 75–81.

Tom, J.A., Reeder, J., Forrest, W.F., Graham, R.R., Hunkapiller, J., Behrens, T.W., and Bhangale, T.R. (2017). Identifying and mitigating batch effects in whole genome sequencing data. BMC Bioinformatics *18*, 351.

Wang, M.-K. (2002). Searching for Qiang culture in the first half of the twentieth century. Inner Asia *4*, 131–148.

Wang, K., Li, M., and Hakonarson, H. (2010). ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. *38*, e164.

Wang, B., Zhang, Y., Zhang, F., Lin, H., Wang, X., Wan, N., Ye, Z., Weng, H., Zhang, L., and Li, X. (2011). On the origin of Tibetans and their genetic basis in adapting high-altitude environments. PLoS One *6*, e17002.

Yang, M.A., Fan, X., Sun, B., Chen, C., Lang, J., Ko, Y.-C., Tsang, C.-h., Chiu, H., Wang, T., Bao, Q., et al. (2020). Ancient DNA indicates human population shifts and admixture in northern and southern China. Science *369*, 282.

Yao, H., Tang, S., Yao, X., Yeh, H., Zhang, W., Xie, Z., Du, Q., Ma, L., Wei, S., and Gong, X. (2017). The genetic admixture in Tibetan-Yi Corridor. Am. J. Phys. Anthropol. *164*, 522–532.

Yi, X., Liang, Y., Huerta-Sánchez, E., Jin, X., Cuo, Z.X.P., Pool, J.E., Xu, X., Jiang, H., Vinckenbosch, N., and Korneliussen, T.S. (2010). Sequencing of 50 human exomes reveals adaptation to high altitude. Science *329*, 75–78.

Yuan, X.-w., and Chen, D. (2011). Soul-sending": the custom of the twice burial under the ethnological perspective— —a study on the definition and reasons of the twice burial based on ethnographies [J]. J. Guangxi Univ. Nationalities (Philosophy Soc. Sci. Edition) *5*, 108–113.

Zhang, M., Yan, S., Pan, W., and Jin, L. (2019). Phylogenetic evidence for Sino-Tibetan origin in northern China in the late neolithic. Nature *569*, 112–115.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
1000 Genomes Project Phase 3	1000 Genomes Project Consortium et al., 2015	https://www.internationalgenome.org/category/phase-3/
The Simons Genome Diversity Project	Mallick et al., 2016	EBI-ENA: PRJEB9586、ERP010710
Tibetan Highlanders Project	Lu et al., 2016	GSA: PRJCA000246 NODE: ND00000013EP
Tibetan-Yi Corridor population variant database	This paper	GVA: GVM000100
Software and algorithms		
SOAPnuke v1.0.0	Chen et al., 2018	https://github.com/BGI-flexlab/SOAPnuke
BWA v0.7.13	Li et al., 2009	http://bio-bwa.sourceforge.net/
SAMtools v1.9	Li et al., 2009	http://samtools.sourceforge.net/
Picard v2.1.0	Broad Institute, 2019	http://broadinstitute.github.io/picard/
GATK v3.8	McKenna et al., 2010	https://software.broadinstitute.org/gatk/
BCFtools v1.9	Li, 2011	https://samtools.github.io/bcftools/
VCFtools v0.1.17	Danecek et al., 2011	http://vcftools.sourceforge.net/
PLINK v1.90	Purcell et al., 2007	https://www.cog-genomics.org/plink2/
ANNOVAR v2018Apr16	Wang et al., 2010	http://annovar.openbioinformatics.org/en/latest/
ANGSD v0.931	Korneliussen et al., 2014	http://www.popgen.dk/angsd/index.php/ANGSD/
BEAGLE v4.0	Browning and Browning, 2007	https://faculty.washington.edu/browning/beagle/b4_0.html
SHAPEIT v2.0	Delaneau et al., 2012	https://mathgen.stats.ox.ac.uk/genetics_software/shapeit/shapeit.html
KING v2.0.1	Manichaikul et al., 2010	http://people.virginia.edu/~wc9c/KING/
EIGENSOFT v4.2	Price et al., 2006	https://data.broadinstitute.org/alkesgroup/EIGENSOFT/
FineSTRUCTURE v2.0	Lawson et al., 2012	http://www.paintmychromosomes.com/
ADMIXTURE v1.3.0	Alexander et al., 2009	http://software.genetics.ucla.edu/admixture/
CLUMPAK v1.1	Kopelman et al., 2015	http://clumpak.tau.ac.il
Ohana v1.0	Cheng et al., 2016	https://github.com/jade-cheng/ohana
EEMS v1.0	Petkova et al., 2016	https://github.com/dipetkov/eems
MSMC v1.0	Schiffels and Durbin, 2014	https://github.com/stschiff/msmc
Admixtools v6.0	Patterson et al., 2012	https://github.com/DReichLab/AdmixTools/

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Shuai Cheng Li (shuaicli@cityu.edu.hk).

Materials availability

This study did not generate new unique reagents.

Data and code availability

The variation data reported in this paper have been deposited in the Genome Variation Map (GVM) in Big Data Center, Beijing Institute of Genomics (BIG), Chinese Academy of Science, under accession numbers GVM000100 at http://bigd.big.ac.cn/gvm/getProjectDetail?project=GVM000100.

This paper does not report original code.

Any additional information required to reanalyze the data reported in this work paper is available from the lead contact upon request.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

We collected the whole blood of 248 participants from 16 populations in the Tibetan-Yi Corridor (TYC) and three Gansu and Inner Mongolia (Table 1). Sampled TYC populations include Tibetans from Sichuan (Tibetan(SC)), Jiarong, Ersu, Pumi from the north (Pumi(N)), Muya, Namuyi, and Yi from Sichuan province, and the Naxi, Pumi, Bai, Han, Lahu, Jino, Jingpo, and Achang from Yunnan province (Figure 1A). The Tibetan(SC), Jiarong, Ersu, Muya, Pumi(N), and Namuyi are officially designated by the People's Republic of China as Tibetan subgroups.

The individuals were recruited using the following criteria: i) with signed written informed consent; ii) random related healthy adults; iii) 1:1 gender-matched in each population; d. 5 mL whole blood was collected by the anticoagulant tube with EDTA. All DNA was extracted within 72 h after blood drawn. All studies and protocols were approved by the Ethics Committee of Xi'an Jiaotong University (approval no. 2016397). Underlying data are submitted to the Management of Human Genetic Resources in China for the record (no. 2020BAT0782).

METHOD DETAILS

DNA extraction, library construction, and sequencing

Venous whole blood samples were collected from 248 participants with 5 mL of vacuum EDTA anticoagulant blood tubes. We extracted a minimum of 2.5 micrograms of DNA from each sample for PCR-based library preparation and sequencing. DNA integrity, purity, and concentration were assessed by agarose gel electrophoresis, the NanoDrop2000 spectrophotometer, and the Qubit 2.0 fluorimeter (Thermo Fisher Scientific). Qualified DNA samples were used for library construction. We process all samples using the same standard Illumina DNA library preparation and sequencing protocol (all library preparation and sequencing took place between the dates of Dec 6th, 2013 and May 7th, 2015 in 20 batches), minimizing the danger that systematic differences in processing could cause artifactual differences among samples. After sequencing, the 38 high coverage samples had 17.24–29.18-fold coverage (median 23.3-fold), and 210 low coverage samples had a range of 4.16–12.14 fold coverage (median 6.7-fold). The average mapping rate from all samples was above 96% and the quality of sequencing data was good (Table S3).

Sequencing was conducted at Beijing Genome Institute on an Illumina HiSeq 2000 platform on 100 bp paired-end reads. We sequenced two unrelated individuals from each population of Han, Hui, Mongolia, and Tibetan in Sichuan (Tibetan(SC)) at a depth of 25× and the other 15 ethnic groups of the TYC with two individuals and 14 individuals at depths of 25× and 5×, respectively (Table 1). The detailed methods and callset generation pipeline were presented in the supplementary materials and Figure 6.

Alignment and variant calling

The SOAPnuke (v1.0.0) (Chen et al., 2018) software was used to remove adaptors contamination and reads with low quality (reads with unknown nucleotides larger than 5%, and reads contain more than 20% nucleotides with quality value $\leq$10). The results of quality control are presented in Table S3. Cleaned reads were mapped to the human reference genome (hg19) with BWA-MEM (v0.7.13) (Li and Durbin, 2009) in paired-end mode. Duplications were detected and removed with Picard (v2.1.0) (Baroud and Steffen, 2005). Base quality score recalibration, indel realignment, variant discovery, and genotype calling were conducted by GATK (DePristo et al., 2011) version 3.8 following the best practice [32]. The genotypes and variants were called by HaplotypeCaller. In addition, ANGSD (v0.931) (Korneliussen et al., 2014) was used to calculate genotype likelihood (GATK model) on a set of high confidence sites. SNVs were annotated against RefSeq with annovar (2018Apr16) (Wang et al., 2010). The specific version and download site of each software tool used was reported in Star Methods.

TYC sequencing data and previously published data combining

Variants from phase 3 of the 1000 Genome project (1KG) (Auton et al., 2015) and the Simons Genome Diversity Project (SGDP) (Mallick et al., 2016), and high-depth sequencing data from Tibetan highlanders Project (THP, including 33 Tibetan highlanders, 30 Hans, and five Sherpas) (Lu et al., 2016) were also incorporated in this study. These dataset resources are summarized in STAR Methods. First, we perform the genotype calling process for TYC and THP populations. Here, we separated samples into two groups: group A containing all samples (n = 316, including 248 TYC and 68 THP samples) and group B only containing high-coverage samples (n = 106, including 38 high-coverage TYC samples and 68 THP samples). Second, we used PLINK 1.9 (Chang et al., 2015) and bcftools 1.9 to merge group A with SGDP and 1KG datasets and group B with SGDP (Figure 6). The sites absented in one file and presented in the other were assigned as missing.

Site quality filtering

We addressed the possible batch effects between different datasets and obtained a set of genomic loci (Figure 1B) deemed reliable for population genetic analysis via widely adopted quality control measures. Those filters were as follows: i) Only bases with an Illumina base quality of at least 20 were included; ii) Only reads with a BWA mapping quality of at least 13 were included; iii) Sites with less than two reads were set to missing value; Sites with missing value in more than 50 samples were filtered out; iv) Hardy-Weinberg proportions: expected genotype frequencies were calculated for each variable site based on allele frequencies based on genotypes called with GATK. Loci with ExcHet = 1 annotated by bcftools 1.9 were filtered out; v) Regions that might have problems with paralogy were excluded based on the accessibility strict mask from 10000 Genomes Projects and 100-mer mappability track in the UCSC Genome Browser. For the set including all samples, we added a filtering condition, "sites absent from dbSNP were removed" to improve the confidence of the sites from low coverage samples. Then, we extracted biallelic SNPs that had a MAF of at least 0.05, pruned the sites for linkage disequilibrium with PLINK 1.9, and finally obtained 947,099 SNPs from all samples and 514,068 SNPs from high coverage samples for subsequent analysis (Table S4).

We evaluated the batch effect of the different datasets using the R package genotypeeval (Tom et al., 2017). Based on the summary metrics computed using genotype calls, we observed no expression batch effects between TYC, THP, SGDP, and 1KG (Figure S7A and S7B).

QUANTIFICATION AND STATISTICAL ANALYSIS

Kinship analysis

We applied KING (v2.0.1) (Manichaikul et al., 2010) to test cryptic relatedness across the TYC individuals (Figure S7C). The kinship coefficients of the three samples were slightly greater than 0.2. After confirming that these three individuals were not related within three generations, we retained these samples.

Haplotype-based analysis

Before pruning, we took genotype likelihoods from GATK at 4,807,702 high-confidence biallelic SNPs pass all filters for phasing and imputation. Firstly, BEAGLE (v4.0) (Browning and Browning, 2007) was run to obtain an initial set of genotypes. SHAPEIT2 (Delaneau et al., 2013) was used to phase the genotype onto haplotypes based on the initial set. Genotypes called by Beagle with a posterior probability greater than 0.995 were fixed as known genotypes. SHAPEIT2 was run with 10 burning-in iterations, 10 pruning iterations. This was followed by 50 sampling iterations were used to estimate the final set of haplotypes.

Principal component analysis (PCA)

To analyze the stratification of populations inhabiting TYC in the context of worldwide and regional people, we performed principal component analyses using the smartPCA program of EIGENSOFT v4.2 (Price et al., 2006) with default parameters. We used genotype data of all samples and genotype likelihood data of high coverage samples, respectively. A series of PCAs, based on genotype data, was performed by selecting close-related individuals based on previous PCA runs to investigate fine-scale population structure. Moreover, we applied fineSTRUCTURE 2.0 (Lawson et al., 2012) to perform linked PCA by using haplotypes of high coverage samples.

Model-based clustering

We performed model-based clustering analysis using the maximum-likelihood approach implemented in ADMIXTURE (v1.3.0) (Alexander et al., 2009) in the context of worldwide populations (LD-independent SNP callset including high coverage samples from TYC, THP, and SGDP) and regional populations (LD-independent SNP callset including all samples from TYC and THP). Under given k components, the highest likelihood ADMIXTURE result over 10 replicates with random seeds were collected. A series of ADMIXTURE analyses, varying K from 2 to 13 for the worldwide population dataset and from 2 to 8 for the regional population dataset, were conducted with the default 5-fold cross-validation setting. We utilized CLUMPAK (v1.1) 38 to identify the best output clusters among replications and aligned clusters with different K values. Moreover, we conducted the population trees for ancestry components from the ADMIXTURE result using Ohana 1.0 (Cheng et al., 2017). We also applied fineSTRUCTURE 2.0 to investigate population structure based on haplotypes of high coverage samples.

Genetic differences between populations (F_{st})

Genetic differences between populations were measured with F_{st}, according to Hudson. This metric is not sensitive to the ratio of sample sizes and does not overestimate F_{st}. We compute Fst with populations from TYC, THP, and 1KG by using EIGENSOFT v4.2 with default parameters. Populations with a sample size smaller than five were discarded.

F_3 statistics and D-statistics analyses

We used the admixture-F_3 statistic f_3(X, Y; test) as implemented in Admixtools (Patterson et al., 2012) to test evidence of the test population (TYC) is derived from the admixture of populations related to X and Y. A significantly negative statistic provides strong evidence of mixture in the test population. The outgroup-F_3 statistic f_3(X, Y; outgroup) were used to estimate shared genetic drift

between X and Y. X and Y are various non-African populations and the Mbuti act as an outgroup. D(X, Y; Z,W) statistic implemented in Admixtools was used to test a tree-like relatedness between four populations. The blocked jackknife was used to assess the statistical significance of the result.

Runs of homozygosity

To study the demographic history of the ancestors of TYC individuals, we investigated runs of homozygosity (ROH) across the genome. We used PLINK 1.9 to screen for runs of homozygous genotypes using sliding windows of 5Mb within all unrelated samples. We defined the minimum length of an ROH to be 500 kb, allowing one heterogeneous call per window.

SFS estimation

Thirty-eight high-coverage TYC individuals were randomly selected for estimating the SFS using the maximum-likelihood method implemented in ANGSD. When we estimated the SFSs with ANGSD, samtools genotype likelihood model was adopted.

Historical population effective size inference

We used the multiple sequentially Markovian coalescent (MSMC) to infer the effective size of the ancestral populations. Only autosomes of high-coverage individuals were used. MSMC studies were performed following the recommendations (Schiffels and Durbin, 2014). We used the mutation rate $1.25e^{-8}$ per base pair per generation for a generation time of 25 years to scale time.

Migration and isolation by distance

We performed geolocation-based analysis using high coverage individuals of TYC and Tibetan highland with EEMS 1.0 (Petkova et al., 2016). Genetic dissimilarities were calculated using the bed2diffs script, and EEMS was conducted with runeems_snps. We ran a burn-in of 1 million iterations for each run, followed by an additional 2 million iterations with posterior samples taken every 1,000 iterations. We assessed the convergence of the Markov chain Monte Carlo by the posterior probability trace plot. The PopGPlot R (v2.7.2) package was used to visualize the data. Additionally, the ancestry coefficients were visualized on the map via interpolation.