Genomic insights into population structure, adaptation, and archaic introgression at the Himalayan–East Asian crossroads

Mengge Wang[2*#], Shuhan Duan[1,2,3#], Qiuxia Sun[2,3#], Yuntao Sun[2,3,4#], Lintao Luo[1,2,3], Yunhui Liu[1,2,3], Renkuan Tang[1*], Libing Yun[3,4*], Chao Liu[5*], Guanglin He[1,2,3*#]

[1]Department of Forensic Medicine, College of Basic Medicine, Chongqing Medical University, Chongqing 400331, China
[2]Institute of Rare Diseases, West China Hospital of Sichuan University, Sichuan University, Chengdu, Sichuan 610000, China
[3]Center for Archaeological Science, Sichuan University, Chengdu, Sichuan 610000, China
[4]West China School of Basic Science & Forensic Medicine, Sichuan University, Chengdu, Sichuan, 610041, China
[5]Anti-Drug Technology Center of Guangdong Province, Guangzhou, Guangdong 510230, China

**Correspondence: Mengge Wang (Menggewang2021@163.com), Renkuan Tang (renktang2012@163.com), Chao Liu (liuchaogzf@163.com), Libing Yun (yunlibing@scu.edu.cn), Guanglin He (guanglinhescu@163.com)*

**Correspondence: Mengge Wang (Menggewang2021@163.com), Renkuan Tang (renktang2012@163.com), Chao Liu (liuchaogzf@163.com), Libing Yun (yunlibing@scu.edu.cn), Guanglin He (guanglinhescu@163.com)*

Abstract
Tibetan-Yi Corridor (TYC) is a crucial agro-pastoral region in the eastern Himalayas, linking Qinghai–Tibet Plateau with the lowlands of East Asia and facilitating human migration for millennia. However, genomic research on TYC populations remains limited, which limits the understanding of their origins and health. We provide genomic data from 1,031 individuals belonging to Austroasiatic and Sino-Tibetan groups, including 147 whole-genome sequences from 13 underrepresented Tibeto-Burman and Austroasiatic communities. Our analysis revealed approximately 3.3 million new genetic variants and four distinct genetic backgrounds within TYC populations. Demographic reconstructions revealed strong genetic connections among Tibeto-Burman groups, Central Plain Sinitic populations, and Yellow River Yangshao farmers, supporting a common origin for Sino-Tibetan speakers. We identified signatures of high-altitude adaptations typical of Tibetans and TYC-specific variants linked to pigmentation and hypoxia responses. Differentiation involves mechanisms such as *HLA-DQB1*, which are related to immune function. Several rare pathogenic variants, like *CYP21A2* and *PRX*, were notably frequent. Variants influencing warfarin sensitivity showed significant variation. Archaic human introgression added complexity, impacting cardiovascular and immune-related genes, which suggests adaptation through ancient human interactions. These findings refine the evolutionary history of TYC populations and underscore the need for broader genomic research to capture regional diversity and inform precision medicine.

Keywords: Tibetan-Yi Corridor; Sino-Tibetan; distinct genetic backgrounds; high-altitude adaptive signatures; archaic introgression; precision medicine

Introduction

Tibetan–Yi Corridor (TYC), located in the eastern Himalayas, connects the Qinghai–Tibet Plateau (QTP) with the lowlands of East Asia, covering many of Sichuan, Gansu, Qinghai, and Yunnan Provinces. Its rugged, river-carved mountains support remarkable ecological, ethnolinguistic, biological, and genetic diversity. As a vital crossroads among the Hexi Corridor, mainland Southeast Asia (SEA), the QTP, and the Central Plain, TYC has long fostered cultural and genetic exchange [1-3]. Its terrain has also helped preserve the isolation of groups such as the Amdo Tibetan (ADT), Lahu, Wa, and Blang. Archaeological sites at Yuanmou (Yunnan) and Piluo (Sichuan) provide evidence of Paleolithic occupation and extensive Neolithic cultural diffusion from northern Gansu and Qinghai [4]. Linguistic and archaeological data reveal complex relationships among Sino-Tibetan sublineages, with competing hypotheses suggesting origins in North China, South China, or northeast India, dating from approximately 4~9 thousand years ago (kya) [1,2]. Some models connect the Sino-Tibetan expansion to Neolithic millet farmers from the Yellow River [1,2], whereas others link early rice agriculture to Austroasiatic dispersal from South China [5]. Interactions between migrating farmers and indigenous hunter-gatherers, along with significant language borrowing

[6], highlight the deep genetic and cultural complexity of TYC populations.

Large-scale genomic resources from Europe [7], Africa [8], Oceania [9], and the Americas [10,11] have greatly improved our understanding of regional evolutionary histories and how evolutionary events affect population migration, disease, and health [12,13]. In contrast, genomic studies of ethnolinguistically diverse Chinese populations are limited, especially for those from TYC, since most autosomal data, Y-chromosome, and mitochondrial DNA focus on Han Chinese individuals and underrepresent minority ethnic groups [14-20]. Located between the highlands of Tibetans, the lowlands of Han, and surrounding indigenous groups, TYC is predominantly inhabited by Tibeto-Burman speakers but also contains linguistically distinct Austroasiatic and Tai-Kadai populations in its eastern and southern parts. The region's cultural, ethnic, linguistic, and biological diversity reflects its history as both a Paleolithic migration route and a pathway for the rapid Neolithic expansion of farmers. Ancient DNA from Yunnan revealed genetic links between early Neolithic populations and older QTP groups, as well as mid-Neolithic communities tied to Austroasiatic origins [21]. Additionally, the introduction of agriculture into eastern TYC, as evidenced by the Gaoshan and Haimenkou relics of the Baodun culture, occurred approximately 4,000 years ago [22]. Although Tibetans on the QTP have been extensively studied [23,24], other Tibeto-Burman-speaking communities within TYC, despite their considerable genetic diversity, have received considerably less attention. Genetic analyses of highland Tibetan and Deng populations revealed shared ancestry with Han Chinese and close links to ancient northern Chinese millet farmers, with divergence times estimated at 15-7.2 kya [25]. Understanding the demographic history of these underrepresented groups is crucial, as migration, gene flow, introgression, and divergence have all influenced their genomes [26]. The eastern and southeastern edges of TYC have undergone repeated and complex cultural and genetic interactions [4]. Archaeological and historical records show that the ancestors of present-day southern Chinese indigenous peoples include the ancient Baipu and Baiyue tribes, along with the ancestors of Austroasiatic speakers [21,27,28]. Ancient DNA studies further revealed that Neolithic farmers from Man Bac in mainland SEA carried ancestry from southern Chinese farmers as well as from deeply divergent indigenous hunter-gatherers [29,30]. Multiple post-Neolithic migration waves involving Austronesian, Hmong-Mien, and Tai-Kadai speakers from South China later reshaped the genetic landscape of mainland SEA and eastern TYC [29-32]. Y-chromosome phylogeny supports the dispersal of the Austroasiatic haplogroup O2a-M95 over the past approximately five kya, which is likely linked to early rice farmers [33,34]. Despite these advances, the detailed genetic structure, demographic history, adaptive traits, and health-related effects of admixture, including archaic introgression from Denisovans and Neanderthals, remain poorly understood for Tibeto-Burman- and Austroasiatic-speaking populations in TYC and their Himalayan neighbors.

Previous genetic studies have explored the population history of the core TYC region from a single perspective. Zhang *et al*. analyzed low-coverage whole-genome sequencing (WGS) data and identified geographic patterns of genetic affinity among TYC populations [3]. Yang *et al*. reported that population stratification in Yunnan, at the southern end of TYC, was shaped by millet-farming Tibeto-Burman groups and rice-farming populations, including Austroasiatic, Tai-Kadai, Hmong-Mien, and Austronesian speakers [28]. Genome-wide single-nucleotide polymorphism (SNP) analyses have also revealed gene flow and biological adaptations that distinguish Yi populations from different geographic regions [35]. Other studies have examined forensic markers, including Short Tandem Repeat, SNP, and Insertion/Deletion variation from autosomal and Y-chromosome data, further emphasizing the region's genetic complexity [18,36,37]. Overall, this molecular evidence reveals the complex evolutionary history of TYC populations. However, many studies remain limited by small sample sizes, low SNP density, narrow population coverage, or limited geographic scope, hindering a comprehensive understanding of migration, admixture, and health-related patterns. Dense sampling across ethnolinguistically diverse groups is essential to clarify the genetic origins and health implications of TYC populations. Exposed to diverse environments, pathogens, and subsistence strategies, these populations offer a unique opportunity to study how evolutionary pressures have shaped genetic architecture and disease susceptibility [38]. Evolutionary research on Tibetan, Sherpa, and Deng populations has revealed strong signals of high-altitude adaptation, including *EPAS1*, *EGLN1*, *HAL-DQB1*, and *GNPAT* [23-25,39,40]. Whether these genes play similar roles in other Tibeto-Burman-speaking groups in TYC and whether additional mechanisms drive rapid adaptation to local environments remain unresolved. Archaic introgression, most notably *EPAS1*, has been shown to contribute to high-altitude adaptation in Tibetans from the QTP [24,41,42]. Physical anthropology studies suggest that the Ziyang people exhibit Neanderthal facial traits, suggesting possible additional gene flow between present-day EAs and archaic humans in TYC [43]. Similarly, the Lahu population carries deeply diverged non-African F lineages, indicating long-term interaction and interbreeding with archaic groups [44]. Despite these findings, the extent of archaic introgression in TYC populations, the demographic history behind these events, and the biological significance of the introgressed segments remain poorly understood.

Here, we performed whole-genome sequencing of seven Tibeto-Burman-speaking populations from the core TYC regions in Sichuan, including previously unreported Qiang, Baima Tibetan (BMT), and ADT groups, as well as four Tibeto-Burman-speaking populations from Yunnan, Bai, Pumi, Hani, and Lahu. We also sampled Wa and Blang, who speak Austroasiatic languages, from Yunnan (**Fig. 1a, b**). In addition, we analyzed genome-wide genotyping data from 884 Sinitic-speaking Han individuals from the Central Plain (CPH), which represents a core ancestral region of Sino-Tibetan speakers, together with published data from the 10K Chinese People Genomic Diversity Project (10K_CPGDP) [45] and publicly available WGS resources from the Human Genome Diversity Project (HGDP) [44] and the 1000 Genomes Project (1KGP) [46] (**Fig. 1a**). This study refines the fine-scale genetic structure of TYC populations, identifies local adaptations, redefines patterns of archaic introgression and their medical relevance, offers insights into population health, and pinpoints archaic-introgressed variants.

Results

Whole-genome sequencing, genotyping, and variant discovery

We collected 1,031 blood and saliva samples from indigenous Sinitic, Tibeto-Burman, and Austroasiatic individuals representing diverse ethnic groups from TYC and neighboring regions (**Fig. 1a, b**). Informed consent was obtained from all participants. Whole-genome sequencing was performed on 147 Tibeto-Burman and Austroasiatic individuals from Sichuan and Yunnan Provinces, achieving mean depths of 12.5× and 91.7% coverage, respectively. After read alignment, variant detection, and strict variant filtering, we identified a set of 13,662,805 biallelic variants with a Ts/Tv ratio of 2.07, along with 1,689,560 small InDels ranging from 1 bp to 50 bp. Among the 13.6 million biallelic SNVs identified, 43% were classified as rare, with a minor allele frequency (MAF) < 0.01. Another 17.5% (2,396,546 SNVs) had an MAF between 0.01 and 0.05. Notably, 24.4% (approximately 3.3 million) of the SNVs were novel, as they were not listed in dbSNP. Most of these novel SNVs were rare, with 89% (2,970,397) having an MAF < 0.01, 7.5% (251,896) between 0.01 and 0.05, and only 3.5% (117,096) being common (**Fig. 1c and Table S1**). To evaluate the biological significance of the novel variants, we annotated them via Variant Effect Predictor tools, with a focus on medical conditions. Intronic and intergenic regions made up 50.22% and 5.57% of the variants, respectively. Only 1.53% of the variants were in coding and splicing regions. Within these regions, missense variants accounted for 15.57%, whereas splicing variants represented 8.77% (**Fig. 1d and Table S2**).

Insight into the fine-scale genetic profiles of diverse TYC groups

To clarify the genetic relationships between TYC populations and global groups, we analyzed WGS data from ethnolinguistically diverse TYC populations, nearby Han Chinese groups, and 185 current worldwide populations. Genetic differences, measured by the fixation index (F_{ST}), were compared between the TYC populations and reference groups (**Table S3**). The TYC populations, located in southwestern China alongside highland Tibetans and lowland Sinitic speakers, presented the closest genetic affinity to EAs ($F_{ST} < 0.06$), followed by SEAs ($F_{ST} < 0.12$). Among the EA groups (**Fig. S1**), ADT, BMT, Bai, and Pumi had the closest genetic ties to the highland Tibetan and Tibeto-Burman populations near the QTP, such as Yi and Naxi ($F_{ST} < 0.02$). In contrast, the Qiang populations presented particularly close genetic proximity to southern Han populations ($F_{ST} = 0.00009$). Moreover, the Lahu, Hani, Wa, and Blang populations presented stronger genetic connections with southern Tai-Kadai groups, such as the CDX and Dai populations in Yunnan ($F_{ST} < 0.03$). These results were supported by principal component analysis (PCA) of the WGS dataset (**Fig. S2a, b**).

To further explore the genetic relationships between the sampled individuals and other Asian populations, we performed additional PCA. Projecting ancient samples onto the first two principal components derived from modern EAs and SEAs revealed that PC1 arranged Tibetans, TYC populations, CPH, southern Han Chinese, and Tai-Kadai/Austroasiatic groups along a north–south gradient. PC2 separated the Tibet–Burman and Sinitic groups along the highland–lowland axis, with TYC populations spread across PC1, highlighting their geographic diversity. Notably, these populations appeared structured into four distinct north–south genetic clines among the EA groups. The leftmost clade, comprising highland Tibetan groups, connected the QTP populations with the ADT populations and Pumi populations from TYC (**Fig. S2c**). The middle cline included BMT, Naxi, and ancient individuals from Neolithic Baodun sites, whereas the lowermost cline included Qiang and southern Han groups. The rightmost cline featured Lahu, Hani, Wa, Blang, mainland SEAs, and Ancient Southern East Asians (ASEA) (**Fig. 1e, S2d, e**). The outgroup-f_3 statistics further reinforced these genetic patterns within the TYC populations, as did the haplotype-based IBD analysis (**Fig. S3**). Haplotype-based fineSTRUCTURE analysis of the Illumina 2240K dataset identified four distinct genetic clusters within the TYC groups (**Fig. 1f**). The first cluster, the TYC high-altitude lineage (THL, blue), included populations from high-altitude regions, such as the core and peripheral areas of the QTP. The second

cluster, the TYC Central Lineage (TCL, red), represented central populations. The third cluster, the TYC Southern Lineage (TSL, green), consisted of Tibeto-Burman speakers such as Hani and Lahu, along with Austroasiatic groups from southern TYC. The final cluster, the TYC Eastern Lineage (TEL, indicated by yellow and purple), encompassed populations with close genetic ties to Sinitic groups, including the Qiang and Bai.

Ancestry makeup and genetic origin of the four ancestrally different TYC lineages

To explore the genetic differences among the four TYC lineages, which are geographically and linguistically close, we performed an unsupervised model-based ADMIXTURE analysis using modern and ancient EA reference populations from the merged dataset, with K values ranging from 2 to 20 (**Fig. 2a**). The model with the lowest cross-validation error, which was considered optimal, identified four ancestral sources (K = 4). The TYC populations mainly shared northern ancestral components, most strongly seen in DevilsCave (green) and Shigatse 2.6K (light blue), along with secondary southern components represented by Yiyang (red). Compared with the other TYC populations, the THL populations, including the ADT, Pumi, and Tibetan populations, were strongly associated with the high-altitude ancient QTP populations (**Fig. 2b**; Wilcoxon, $P < 0.05$). TCL populations, such as BMT, Yi, and Naxi, presented relatively high levels of northern EA ancestry, like Neolithic Gaoshan individuals. Qiang and Bai, which fall under TEL, share genetic components with Neolithic Yangshao/Longshan-related groups, especially China_YR_MN and China_YR_LN. In contrast, the TSL populations presented a greater proportion of southern modern and ancient ancestry, setting them apart from the other TYC groups (**Fig. 2b**; Wilcoxon, $P < 0.05$). These model-based findings suggest that the genetic structure of TYC populations results from a mixture of high-altitude, northern, and southern ancestral components, with different proportions across the THL, TCL, TSL, and TEL groups (**Fig. 2a, b**). Additionally, the high-resolution WGS dataset revealed unique demographic histories for the THL and TSL groups, as indicated by distinct ancestral components specific to Wa, Hani, Lahu, ADT, and Pumi (**Fig. S4**).

We further tested potentially differentiated admixtures contributing to the observed genetic differentiation and performed symmetric f_4-statistics in the form of f_4(TYC1, TYC2; modern reference populations, Mbuti) based on the 2240K_HO dataset. Positive Z-scores were observed, indicating that THL and TCL shared more alleles with Tibeto-Burman speakers, such as those previously reported by Ü-Tsang, Kham, and Ando Tibetans and Ancient Northern East Asians (ANEAs), than did TSL and TEL (**Fig. 2c, d**). Notably, TCL received additional gene flow events from Tai-Kadai/Hmong-Mien/Austronesian-speaking people and the ASEA relative to THL (**Fig. 2c, d and Table S4**). Similarly, compared with the other TYC populations, the TSL population received additional gene flow from modern Austroasiatic speakers and ancient individuals from mainland SEA (**Fig. 2c, d and Table S5**). These results further support the existence of population substructures and the genetic influence of different genetic ancestries among genetically distinct TYC populations.

We then investigated whether the genetically distinct TYC populations originated from a common source. Using the qpWave method, the p_rank0 estimates showed no significant differences, suggesting genetic similarity between Neolithic Gaoshancheng individuals and modern TCL individuals when nearby ancient samples were included (**Fig. S5**). We further examined the genetic uniformity or diversity among these populations, as well as modern Sino-Tibetan groups and ancient QTB populations. qpWave analysis revealed geography-related population substructures, in which geographically and linguistically close populations possessed heterogeneity (**Fig. S6**). Additional analyses incorporating ancient ancestors, including Neolithic Yellow River millet farmers, Neolithic rice-farming groups from Fujian and Guangxi, and Paleolithic hunter-gatherers such as Zongri5.1k and Hòabìnhian, were performed. These analyses, which were based on the ADMIXTURE results and qpAdm models, helped define the genetic components of the current THL, TSL, and TEL populations. Most Tibeto-Burman and Austroasiatic groups inherited ancestry from Middle Neolithic Yangshao-related YRB farmers, with minor contributions from the ASEA (**Fig. 2e**). The ancestry and admixture patterns in THL, TCL, and TEL closely matched the model fitted for CPH (China_YR_MN 67%, China_SEastAsia_Island 33%). When two-way models for the post-late Neolithic period were used, the highest level of Qijia-related (~96%) ancestry was found in the THL gene pool. Conversely, TSL had less YRB-related ancestry and more Yangtze rice farmers than the other TYC lineages did (**Fig. S7a, Table S7**). This admixture variability was further supported by qpGraph analysis (**Fig. S7b-d**) and f_4-statistics (**Table S6**). We further used qpAdm models to clarify the genetic makeup of other Sino-Tibetan and ancient QTB populations, confirming their primary ancestry from millet farmers and revealing diverse admixed gene pools (**Fig. S8**). These findings suggest that TYC Tibeto-Burman-speaking populations share a common origin with ancestors linked to the CPH and are associated with millet farmers from the YRB in northern China. Furthermore, the four lineages experienced varying levels of genetic input from different millet-farming groups since the late Neolithic.

Time depth and demographic mode of the TYC population separation

To estimate the timing of genetic divergence between the TYC populations and Sinitic people, we applied MSMC2, which assumes an instantaneous split with no subsequent gene flow between populations. The split time was defined as the point at which the estimated coalescence rate between two populations was half that within a single population. Using a mutation rate of 1.25×10^{-8} per base pair per generation and a generation time of 29 years, the midpoint estimates indicated a split between Tibeto-Burman and CHS at ~13-14 kya, between Austroasiatic and CHS at ~14 kya, and between Tibeto-Burman and Austroasiatic populations at ~14-16 kya (**Fig. 2f**). Divergence within the TYC populations revealed the earliest pre-Neolithic split between the TSL and the other TYC groups at approximately 15-21 kya, including splits between the TSL and TCL at ~21 kya, between the TSL and THL at ~16 kya, and between the TSL and TEL at ~15 kya. Pairwise RCCR estimates placed the split between THL and TCL at ~13 kya, with a later split from TEL at ~12 kya (**Fig. S9**). These results support a model in which proto-TYC Tibeto-Burman speakers first diverged from proto-Austroasiatic speakers during the Upper Paleolithic, subsequently splitting from the common ancestor of proto-Sino-Tibetan speakers approximately 13-14 kya. This coincides with previous estimates of Tibetan–Han divergence ranging from ~9-15 kya. Additionally, the common ancestors of the Tibeto-Burman groups in TYC further diverged into northern, central, and southern branches, approximately 10-13 kya, reinforcing the idea that TYC was a genetic melting pot.

We also applied nucleotide diversity-based Pi and site frequency spectrum-based SMC++, utilizing data from a larger set of unphased genomes to estimate effective population sizes. Most TEL and TCL populations, including the Qiang, Bai, and Yi populations, which share significant agriculture-related ancestral components, exhibited substantial population growth over the past 11 kya. This trend aligns with a broader pattern of human prehistory, where the effective population size increased as agricultural groups expanded (**Fig. 2g and Fig. S10a**). In contrast, THL and TSL populations, such as ADT, Pumi, Lahu, Wa, and Blang, exhibited modest growth while maintaining high levels of runs of homozygosity (ROH) (**Fig. 2g and Fig. S10b**) and long segments of identical by descent (IBD) between individuals (**Fig. S3b**). These patterns suggest long-term low but stable effective population sizes, indicating recent population isolation, which contrasts with the expansive growth observed in the TEL and TCL populations.

To further investigate potential historical admixture events in the TYC populations, we applied the haplotype-based method fastGLOBETROTTER, which uses 17 genetically distinct populations as proxies for admixture sources (see Methods). Eight out of the 13 TEL and TSL groups, including Qiang, Bai, Hani, Lahu, Wa, and Blang, could be modeled as mixtures of northern and southern ancestries. Northern ancestry, represented by Han individuals from Shaanxi Province, contributed more than 60% to the Tibeto-Burman-speaking groups, including the Qiang, Bai, and Hani. In contrast, southern ancestry, exemplified by Cambodians, accounted for a significant portion of Austroasiatic groups such as Wa (63%) and Blang (66%). The estimated dates for "one-date" admixture events ranged from 43~48 ± 8~15 generations ago for the TSL groups (587~732 Common Era, CE) and ~12~21 ± 1~6 generations ago (1370~1631 CE) for the Qiang within TEL. The ADT, Pumi, and BMT populations could be modeled as admixtures between Tibetan highlanders and lowland Han, mainly among THL and TCL (**Fig. S11a and Table S8**). The 587-732 CE period saw the TYC as a strategic frontier linking the QTP, Hexi Corridor, and East Asian lowlands, with Tang-Tubo alliances, Nanzhao expansion, and Tea-Horse Road trade driving gene flow, supported by Tibetan-period carvings in Shiqu. LD-based MALDER dating revealed admixture events 14~24 ± 1~10 generations ago (~406-696 before present), reflecting gene flow between northern and southern EA sources at the CPH and TYC (**Fig. S11b**). The 1283-1573 CE period featured Han southward migration, Tusi-mediated intermarriage, Mongol incursions, and Nuosu expansion, fostering multi-ethnic admixture. These results reveal that prehistoric divergence, agricultural dispersal, and recurrent historical gene flow jointly shaped the TYC genomic landscape.

Identifying signatures of natural selection

Adaptive selection of genomic loci in response to extreme environments, pathogen exposure, or shifts in dietary habits, such as high altitudes with low oxygen pressure (*EPAS1*, *EGLN1*, and *HLA-DQB1*) and intense ultraviolet radiation (*GNPAT*), has been well documented in Tibeto-Burman populations, particularly Tibetans living on the QTP. To investigate potential population-specific adaptations in TYC populations, we conducted a genome-wide scan via population branch statistics (PBS), with Chinese Han individuals from Beijing (CHB) and Europeans from HGDP (EUR) used as ingroup and outgroup reference populations, respectively. SNPs in the top 0.1% of the PBS scores were considered candidates for natural selection, with only protein-coding genes treated as adaptive candidates (**Fig. 3a-b; Table S10**). We identified 17 candidate genes, 16 of which are newly reported in this study. Among these,

OPCML, previously linked to high-altitude adaptation in the large-scale Tibetan genome sequencing project [24], also emerged as a signal in TYC (**Fig. 3a**). Gene Ontology (GO) analysis revealed that these genes were enriched in pathways related to adaptive physiological traits in TYC individuals, particularly multicellular organismal processes (GO:0032501) and responses to stimuli (GO:0050896) (**Fig. S12a**). Notably, our analysis revealed genetic variants linked to genome-wide association study (GWAS) signals, including blood protein level-related *PKP2*-rs11052294, total cholesterol level-related *AKT1*-rs2494742, and longevity-related *CFAP69*-rs9691522 variants, in at least two or more population lineages (**Fig. 3c; Table S10**). Allele frequency spectrum analysis for these loci revealed higher frequencies of derived alleles in EAs than in other continental populations (**Fig. S12b-d**).

To identify population-specific genetic signals in THL and Tibetans, two genetically distinct but altitude-similar groups, we conducted a genome-wide scan via the $PBS_{THL\text{-}CHB\text{-}EUR}$ trio model. This analysis yielded 563 candidate genes, including well-known high-altitude adaptation genes previously identified in Tibetans[23-25]. Among these variants were the *EPAS1*-rs76347095 (PBS = 0.25) and *EGLN1*-rs58696423 (PBS = 0.25) variants, which have been shown to help maintain lower hemoglobin concentrations, thus preventing polycythemia at high altitudes [40,47]. Notably, we identified a set of genes showing enriched signals of positive selection related to high-altitude adaptation (**Tables S11, 12**). Among them, we discovered a novel missense variant, rs885479, in the *MC1R* gene (**Fig. 3d**). *MC1R*, known for its role in human pigmentation variation [48], was functionally predicted based on a SIFT score of < 0.05. Extended haplotype homozygosity (EHH) analysis revealed extended homozygosity for this variant, and its higher frequency in THL populations than in other EA populations suggests that *MC1R*-rs885479-A may have been a target of selection related to skin pigmentation (**Fig. 3e, Fig.S12e, Fig. S13**). We further analyzed the expression patterns of this gene using GTEx and found high expression in the ovary, thyroid, brain (cerebellar hemisphere and cerebellum), pituitary, and testis (**Fig. S14**). QTP receives intense ultraviolet radiation, including at high altitudes, from TYC. Yang *et al*. recently examined how ultraviolet selection pressures impact human skin pigmentation phenotyping and reported that highland Tibetans had significantly darker baseline skin color than Han Chinese individuals did [39]. Yang *et al*. also reported that *GNPAT*-related signatures contributed to this pigmentation variation, along with our identified potential targets.

Different detection methods capture distinct types and timescales of selection. To assess this, we compared candidate signatures exhibiting outlier PBS values with the haplotype-based integrated haplotype homozygosity score (iHS), which identifies more recent sweeps. The genome was divided into 20 kb windows, and windows with |iHS| values greater than 2 were considered candidate SNPs. Regions where more than 20% of the SNPs met these criteria were selected as adaptive regions. A region under positive selection was identified in THL, containing the signal-induced proliferation-associated 1 like 3 (***SIPA1L3***) gene, which encodes a GTPase-activating protein (**Table S12**). Variants within this region, including rs10421944 (iHS = 2.73), rs2384779 (iHS = 2.54), and rs138460641 (iHS = 2.50), presented strong long-range haplotypes indicative of recent positive selection, which was supported by significant iHS values (**Fig. S12f-h and S15**). We further analyzed gene expression levels within the GTEx database, focusing on the various genotypes of these key mutations (**Fig. S16**). Our findings reveal a strong correlation between selection signals and gene expression in tissues involved in lung, cultured fibroblasts, testis, spleen, colon-transverse, prostate, brain-cerebellar hemisphere, brain-cerebellum, minor salivary gland, skin (not sun-exposed, suprapubic), skin (sun-exposed, lower leg), small intestine (terminal ileum), thyroid, and EBV-transformed lymphocytes. These results independently support the potential role of *SIPA1L3* in the cardiopulmonary functions and response to hypoxia.

To investigate whether geographically diverse TYC populations share similar adaptation mechanisms, we examined similarities and differences in altitude-related selective signals between different TYC groups. These signals, identified in middle-altitude TCL populations, were absent in low-latitude TSL populations (**Table S9**), suggesting that isolated TSLs in forest-grassland ecotones underwent distinct local adaptations. Notably, we detected strong signatures of natural selection in *HLA-DQB1* (PBS = 0.23, iHS = 2.71), *HLA-DQA1* (PBS = 0.50, iHS = 3.09), and *HLA-DRB1* (PBS = 0.21, iHS = 2.44) via the $PBS_{TSL\text{-}CHB\text{-}EUR}$ trio model and haplotype-based iHS approach. These genes, which are part of the major histocompatibility complex class II, have been linked to several autoimmune diseases, including type 1 diabetes, celiac disease, and multiple sclerosis (GO:0002399). We identified *HLA-DQB1*-rs1049083 as a potential target of natural selection, exhibiting a strong EHH signal. Linkage disequilibrium (LD) was observed with upstream SNPs (rs3998178, rs1766) and downstream SNPs (rs41264352, rs4993140), all of which were within the top 0.01% of the PBS values (**Fig. 3f**). These four variants, in LD with *HLA-DQB1*-rs1049083, have been associated with immunometabolism and the body shape index, as noted in the GWAS catalog (**Fig. 3g and Fig. S12i**). We confirmed the *HLA-DQB1*-rs1049083 signal and its linkage disequilibrium (LD) with immunometabolism-related SNPs in an additional TYC anthropological cohort (**Fig. S17**). We also explored biological functions via publicly

available eQTL datasets and reported that rs1049083 is associated with differences in *HLA-DQB1* expression in whole blood, Lung, and the spleen (FDR < 0.05), suggesting a potential immunoregulatory role (**Fig. S18**). Other immune-related adaptive genes, such as *CD58*, *PELI1*, and *NLRP1*, were also identified. These genes were enriched in pathways related to autoimmune diseases, including interleukin-8 production regulation (GO:0032677) and inflammasome-mediated signaling (GO:0141084). Additionally, metabolic variants under positive selection, such as *KMT2A*-rs514924 and *HLA-DRB5*-rs11967891, were found to be associated with HDL cholesterol levels, which is consistent with GWAS catalog findings.

We further examined the selection profiles of low-altitude populations, focusing on Han groups from different regions, including the Central Plain, Fujian, Shaanxi, and Sichuan, to identify unique selection signals across central, southeastern, northwestern, and southwestern China. Using the PBS approach, we identified 30 shared genes, including *DYM* and *FAM129B*, involved in skeletal system development, as well as *PSPC1* and *ENOX1*, which are linked to rhythmic processes. Additionally, 21 genes under natural selection, particularly those related to metabolism (*ABCC11*, *DPYD*, and *ZNF536*), were found in CPHs from Shaanxi and Han from Sichuan; these genes were not detected in southern Han from Fujian (**Fig. S19, Table S13**). These findings suggest that Han Chinese populations from agro-pastoral transitional zones have distinct adaptive genetic foundations compared with those from the southeastern coastal regions of China.

The context for medically relevant variants

To provide population-specific insights into medically relevant genetic variation in TYC populations, we annotated our genomic dataset via the American College of Medical Genetics and Genomics Secondary Findings gene list (ACMG v3.2) [49]. Reportable variants were identified in only three individuals, each carrying a singleton variant in *TTR*-rs376790729-C-A (linked to dominantly inherited transthyretin amyloidosis) or the doubleton variant *MUTYH*-rs121918095-G-A (associated with recessively inherited polyposis) (**Fig. S20a and Table S14**). In contrast, 124 individuals (85%) in the TYC variant dataset carried at least one variant classified as "pathogenic" (P, level 5) or "likely pathogenic" (LP, level 4) according to the ClinVar Database. Each person carried a median of two alleles (range: 1-5) (**Fig. 4a**). Among the 58 unique variants annotated as P/LP, approximately 69% (40/58) had a derived allele frequency less than 0.05. Among these, four variants (6.9%) presented derived allele frequencies above 0.05 across all population groups in the PGG. Han [50] (**Fig. 4b**). Notably, we identified some pathogenetic variants in *CYP21A2, NPHS2,* and *PRX*, with carrier frequencies exceeding 5% for genes associated with classic congenital adrenal hyperplasia due to 21-hydroxylase deficiency, peripheral neuropathy, and urological diseases, respectively, in TYC populations (**Fig. 4c, Fig. S20b, c and Fig. S21a, b; Table S15**). The site frequency spectra of the predicted damaging variants aligned with expectations of purifying selection while also revealing many shared and common P/LP variants, underscoring their potential relevance to variant curation efforts.

Understanding pharmacogenomic diversity, which refers to the variation in allele frequencies that influences individual responses to medication, has significant clinical implications beyond genetic disease risk. To assess this landscape, we identified pharmacogenetic alleles of genes listed in the CPIC drug–gene pair list via PharmGKB level 1 evidence. Every individual in the TYC variant dataset (100%) carried at least one actionable pharmacogenetic variant in 23 genes associated with high-confidence gene–drug interactions, with a median of eight findings per individual (**Table S16**). This high frequency was driven primarily by the *VKORC1*-rs9923231-C-T allele, decreasing the expression of *VKORC1* protein expression and increasing the sensitivity of warfarin, which is present in 91.8% of individuals, affects sensitivity to the anticoagulant warfarin, and is known for its prevalence among Asians, suggesting a reduced dosage [51,52] (**Fig. 4d**). In genetically distinct TYC populations, however, the allele was observed at relatively low frequencies. Additionally, actionable pharmacophenotypes were identified in *CYP2C19*, a gene crucial for the metabolism of commonly used medications, including the antiplatelet agent clopidogrel, proton pump inhibitors, and selective serotonin reuptake inhibitors. Notably, the frequencies of pharmacogenetic variants varied across different genetic backgrounds, with the TSL population showing the highest prevalence (**Fig. 4d and Fig. S21c,d**).

Archaic introgression and the biological functions of high-confidence introgressed fragments

Eurasians exhibit introgressed sequences from both Denisovans and Neanderthals [53,54], although previous studies have focused mainly on Tibetans from the QTP [41,55]. To identify high-confidence sequences likely introgressed from these archaic hominins, we applied IBDmix on an individual basis, intersected the results with introgression segments inferred from SPrime to identify high-confidence introgressed Neanderthal fragments and used SPrime to identify high-confidence Denisovan fragments (see Methods; **Fig. 5a**). In the TYC, after strict quality control, individuals carried approximately 11.46 Mb of

Neanderthal-derived sequences and 4.84 Mb of Denisovan-derived sequences. We applied the same method to examine archaic introgression patterns in highland Tibetan and lowland Han Chinese populations. The analysis revealed similar trends of introgression and identified region-specific introgressed genes. These findings highlight the importance of further exploring the gaps in understanding the archaic introgression in TYC populations (Fig. S22). In total, 413 segments likely introgressed from Neanderthals and 40 from Denisovans were identified in TYC people, spanning approximately 149.42 Mb and 5.97 Mb of the genome, respectively (**Fig. 5b**). The match rate distribution from SPrime revealed a higher rate of Neanderthal introgression than Denisovan introgression across all the TYC populations. Additionally, two distinct peaks were observed, representing genomic segments with low affinity for Neanderthals and higher affinity for Denisovans (**Fig. 5c**). We further conducted two sets of f_4-statistics in the form of f_4(Neanderthals, Denisovan; TYC populations, Mabuti) (Figs. S23-24) and of f_4(TYC1, TYC2; Neanderthals/Denisovan, Mabuti) (Table S17), which confirmed that TYC populations shared more alleles with Neanderthals compared to Denisovan. This bimodal distribution of Denisovan match rates supports the hypothesis of two pulses of admixture with Denisovan-like archaic humans in EA [54] (**Fig. S25a**). We identified the well-characterized Tibetan-*EPAS1* haplotype in the TYC populations, which influences cardiopulmonary functions in the lung, tibial artery, and coronary artery, as observed in the GTE database (Fig. S26). Additionally, we annotated the high expression of CHRND in the Musculoskeletal system (Fig. S27). Notably, haplotypes in northern TYC populations were much closer to the Denisovan haplotype than to any modern human haplotype, particularly when considering the five-SNP Tibetan–Denisovan haplotype (**Fig. S25b**), suggesting that TYC individuals shared similar EA-specific Denisovan introgression events, as seen in Tibetans.

We analyzed the potential phenotypic effects of the identified introgressed sequences via GWAS summary statistics derived from the integrated IBDmix and SPrime results. A total of 239 archaic segments were associated with various phenotypes, with 215 from Neanderthals and 24 from Denisovans. Among these, 104 Neanderthal-inherited segments and 13 Denisovan-inherited segments were linked to genes associated with body mass index (BMI) and height in TYC individuals (**Fig. S28a and Table S18**). Additionally, a Denisovan-like adaptive haplotype at *DIS3L2*, which is associated with lung function, was identified. Three Neanderthal-like missense variants, including *CCDC138*-rs1724184, *CEP83*-rs2271979, and *IARS*-rs471478, were associated with birth weight, pulse pressure, and the COL1A1/OMD protein ratio, respectively, and were highly prevalent in TYC individuals (**Fig. S28b and Table S19**). Distinct adaptive Denisovan-introgressed haplotypes were also identified in differentiated TYC populations. Specifically, THL-specific *EPAS1*-rs7582701 was associated with IgG glycosylation (**Fig. 5d**), which regulates stress erythropoiesis during hypoxia [56]. Additionally, the TSL-specific *CHRND*-rs2245601 signal was linked to the acetylcholine receptor signaling pathway (GO:0095500), which modulates metabolic levels (**Fig. 5e and Table S20**). In TEL individuals, the *CXADR*-rs186722413 adaptive introgression was enriched in cardiac muscle cell adhesion (GO:0086042) (**Fig. 5f and Table S19**). Further analysis revealed the Neanderthal-introgressed adaptation profile across genetically distinct TYC populations. Segments at *VPS11*, *EXOC8*, and *USP36* were associated with cardiovascular function in the THL, TCL, and TEL groups, whereas haplotypes at *UBAP2* and *ARCN1* were implicated in metabolic regulation in the TSL group (**Fig. 5g, h, and Table S21**). These findings suggest that archaic introgression contributes to environmental adaptation and that differentiated adaptive introgression has intensified population differentiation within TYC populations.

Discussion

Fine-scale demographical history supports the northern China origin of the Sino-Tibetan people

Although the number of human genomes sequenced in medically motivated genetic studies has rapidly expanded to hundreds of thousands [7,57-59], anthropologically informed sampling aimed at characterizing human diversity, particularly among ethnolinguistically diverse populations in the eastern regions of QTP, remains underrepresented (**Fig. 1a**). Recent efforts incorporating such sampling strategies have yielded valuable insights into patterns of human genetic diversity and evolutionary history [18,19,44,45] and are critical for developing robust haplotype reference panels [57] and genetic research databases [8,11,51]. In this study, 1,031 genomes from the agro-pastoral transitional zone were analyzed, providing a comprehensive genetic profile of these diverse Chinese populations. Four distinct genetic backgrounds of the TYC Tibeto-Burman and Austroasiatic-speaking groups were identified via fineSTRUCTURE analysis (**Fig. 1**). The TYC served as a crucial corridor for interaction between northern and southern ethnic groups in the eastern regions of the QTP, occupying a unique position for population contact. Genetic differentiation along a north–south cline within the TYC was observed in the first principal component, which included Sichuan Tibeto-Burman populations, Yunnan Tibeto-Burman speakers, and Yunnan Austroasiatic speakers (**Fig. 1e**). Genetic affinity and gene flow were further supported by allele-based f_3/f_4 statistics and haplotype-based IBD analysis. Additionally, language- and geography-associated

genetic substructures were identified among Baima, Ü-Tsang, Amdo, and Kham Tibetans (**Fig. S2c**). The genetic connection between the Austroasiatic and Tibeto-Burman groups in Yunnan indicates significant interactions among nearby populations that are linguistically diverse. These debated patterns, involving language mixing, have also been recently seen in other TYC groups and the Tujia people [3,6]. The TYC region was found to contain the most genetically diverse populations in China, reflecting its complex biological roots. Notably, the Qiang population sampled in northern Sichuan presented a distinct genetic component compared with previously published data from Qiang individuals in western Sichuan [60] (**Fig. S4b**). These findings underscore the crucial role of anthropological sampling in bridging gaps in our understanding of human genetic diversity.

Compared with the Central Plain, which is recognized as the cradle of the Sino-Tibetan language family and Chinese civilization, archaeological materials from sites unearthed in TYC, such as Gaoshan, Haimenkou, and Jinsha, reveal distinct ethnic characteristics based on craftsmanship. These findings provide compelling evidence that indigenous populations in TYC were distinct from inland populations during the prehistoric period and that ancient DNA has the potential to elucidate their genetic origins and differential evolutionary processes (**Fig. 2**). The well-fitted models using qpAdm demonstrated that both CPH Sinitic populations and genetically distinct Tibeto-Burman speakers inhabiting TYC shared a common origin from the YRB during the Middle Neolithic (**Fig. 2e**), which is consistent with previous evidence from linguistic, archaeological, and ancient DNA studies [1,2,22,61]. Additionally, this population genetic work systematically revealed multiple genetic ancestries among genetically differentiated TYC populations since the late Neolithic, with Gaoshan and Haimenkou serving as the best proxy ancestry for the Baodun culture of BMT, Naxi, and Yi, whereas Lajia from the Qijia culture represents the primary proxy ancestry for ADT and Pumi, as supported by f_4-statistics and qpAdm analyses (**Fig. S6 and Table S8**). These findings align with archaeological evidence indicating two distinct southward migrations of TYC populations [62]. Moreover, the divergence times estimated via MSMC2 suggest that the TYC and CHS populations began to diverge at approximately 13-14 kya (**Fig. 2f**), with the earliest separation occurring between Tibeto-Burman and Sinitic speakers, indicating the origin of the proto-TYC populations in the upper YRB, related to millet farmers, which later gave rise to distinct genetic backgrounds in different groups. Additionally, by applying LD decay-based MALDER and haplotype-based fastGLOBETROTTER methodologies (**Table S9**), the north–south migratory movements of proto-TSL/TCL and the eastward migration of highland-related ancestry from the QTP into northern TYC for THL during the historical period were reconstructed. These patterns align with ethnological records (**Fig. S9**). These results fill significant gaps in the genetic understanding of the parallel migration routes of Tibeto-Burman speakers through TYC and the demographic history of populations in this region. These findings underscore the importance of leveraging multi-population WGS resources to disentangle the complex formation patterns of TYC populations. Our findings of close genetic links between TYC populations and northern Chinese millet farmers support the hypothesis that Sino-Tibetan languages originate from north China, as suggested by the reconstructed language phylogeny [1,2]. Additionally, detailed population substructure aligned with geography and linguistic groups matches the structure of the early language phylogeny, reflecting dispersal from northern China with different levels of divergence.

Selection and introgression signatures of the Himalayan–East Asian crossroad populations

The TYC was identified as a geographical transition zone between the high-altitude QTP and the lowland EA, characterized by a topography that slopes from northwest to southeast. This positioning enabled the identification of both shared and population-specific genetic adaptation signatures associated with population differentiation. Shared candidate SNPs in the top 0.1% of scores were mapped to 17 genes (**Fig. 3a**) and GO pathway analysis revealed enrichment in multicellular organismal process (GO:0032501) and response to stimulus (GO:0050896) genes (**Table S12**). Among these genes, the *OPCML* gene, previously associated with high-altitude adaptation, was identified via low-coverage WGS resources [24]. These findings suggest that the TYC population exhibited a shared pattern of adaptation to mid- to high-altitude environments. Additionally, different adaptive genes and functional enrichment categories indicated that the unique living environments and habits of the THL and TSL populations are likely shaped by specific adaptive evolution. A total of 563 candidate high-altitude adaptation genes were identified (**Table S10**), including the well-characterized *EPAS1* and *EGLN1* genes. Allele frequency-based *MC1R* and haplotype-based *SIPA1L3* were detected in THL, with *MC1R* associated with skin pigmentation, suggesting adaptation to high-altitude environments with elevated ultraviolet radiation (**Fig. 3d**). *SIPA1L3*, which encodes a GTPase-activating protein involved in HIF-1 activation, further supports this high-altitude adaptation, as the small GTPase Rac1 is known to be activated under hypoxic conditions, promoting HIF-1 protein expression and transcriptional activity [63]. Previous analyses have indicated that the Deng people possess a unique genetic basis for high-altitude adaptation involving the *HLA-DQB1* gene [25]. This gene was also identified in TSL populations inhabiting the hilly and forested

regions of southern TYC. A novel *HLA-DQB1*-rs1049083 signal, which displays strong evidence of positive selection (**Fig. 3f**), was observed, with robust linkage disequilibrium to SNPs with immunometabolism-related functions annotated in the GWAS Catalog (**Fig. 3g**). These findings suggest that the genetic basis of adaptation in TSL reflects a population-specific response to dietary and pathogen pressures rather than convergent adaptation to high-altitude environments. Leveraging large-scale genomic resources from the CPH populations, 30 high-quality genes associated with skeletal system development and rhythmic processes were identified (**Table S14**). These genes were shared with genomes from the northwest, southeast, and southwest Han populations (**Fig. S13**), as documented in the 10K_CPGDP [45]. Moreover, distinct adaptive genetic bases associated with metabolic pathways were observed in Sinitic-speaking populations from the agro-pastoral transitional zone of China, including CPH, compared with southeastern coastal Han populations.

This study investigated the genomic and phenotypic impacts of Denisovan and Neanderthal DNA within four distinct genetic backgrounds of TYC populations, each of which has undergone unique environmental adaptations. Using IBDmix and SPrime, 453 introgressed segments were identified, with 413 likely derived from Neanderthals and 40 from Denisovans (**Fig. 5a**). Evidence for two waves of Denisovan admixture and one wave of Neanderthal introgression was observed, which was supported by match rate computations (**Fig. 5c, Fig. S16a**) and aligned with prior findings in EA populations [54,64]. Introgressed Denisovan sequences at the *EPAS1* locus have previously been linked to high-altitude adaptation in Tibetans [41]. Similarly, northern TYC populations, including Qiang, ADT, and BMT, presented a closer relationship between the *EPAS1* haplotype and the Denisovan haplotype than other modern populations did (**Fig. S16b**). Two SNPs at the *EPAS1* locus (rs10193827 and rs7582701) were identified as contributing to THL and TCL adaptations to hypoxia (**Fig. S17b**). In addition to specific cases such as *EPAS1*, the broader impacts of archaic introgression on human phenotypes remain poorly understood. Adaptive Denisovan-introgressed strains, such as the TSL-specific *CHRND*-rs2245601, associated with metabolism (**Fig. 5e**), and TEL-specific *CXADR*-rs186722413, linked to the cardiovascular system (**Fig. 5f**), have been reported. Additionally, adaptive Neanderthal-derived segments at *VPS11*, *EXOC8*, and *USP36* were found to impact the cardiovascular system in THL, TCL, and TEL, whereas segments at *UBAP2* and *ARCN1* influenced the metabolic system in TSL (**Fig. 6g-h, Table S21**). These findings demonstrate that archaic introgression facilitates environmental adaptation in modern humans and that differential adaptive introgression contributes to the genomic diversity of TYC populations. Moreover, the population specificity of significant introgressed variants in EAs was highlighted, suggesting that archaic variant–phenotype associations might be overlooked when relying solely on European data. Further research on genetically diverse EA populations will be necessary to understand these associations fully.

Evolutionary insights and medical implications of under-representative TYC populations

The persistent underrepresentation of TYC populations in human genetic studies has limited their inclusion in global genomic datasets and hindered comprehensive assessments of genetic variation [19,36,58]. In our dbSNP-based catalog, we identified approximately 13.6 million biallelic SNPs, including 3.3 million novel variants, with 89% of these occurring at rare frequencies. Clinically important variants, such as dominant and recessive ACMG risk alleles and P/LP variants in ClinVar, were frequently observed, supporting previous evidence that disease burden varies among populations and is often misestimated when primarily based on European cohorts. These findings emphasize the need to include diverse populations to improve global disease risk assessments. *CYP21A2*-rs6467-C-G, linked to congenital adrenal 21-hydroxylase deficiency, was found at >5% frequency in TYC but remained rare in gnomAD East Asians, indicating a higher prevalence possibly due to improved variant curation [65]. Similarly, *PRX*-rs3814290-C-T, rare worldwide, was common among TSL individuals, reflecting local genetic structure. These patterns show how alleles causing Mendelian diseases can be more frequent in underrepresented groups, sometimes due to misclassification in databases like ClinVar or the founder effect of geographically isolated populations [12]. The SEA3K dataset uncovered ten variants common in MSEA but rare elsewhere, many likely misclassified as pathogenic [66]. Likewise, H3Africa data revealed that over 20% (54 out of 262) of the predicted ClinVar pathogenic variants exceeded 0.05% frequency in certain African groups [8]. Overall, these results underscore the importance of diverse genomic resources for accurate variant annotation, filtering the annotated risk variants, and preventing misleading disease risk predictions. Elevated frequencies of variants, such as *GJB2*-related deafness variants in southern China [67] and *HBA2* thalassemia pathogenic variants in Southeast Asians [66], may result from founder effects or local balancing adaptation driven by specific evolutionary forces [12]. Similar patterns appeared in disease-related alleles, such as *G6PD* hemolysis risk alleles (which confer protection against severe malaria), HBB-rs334 (a common sickle cell mutation), *APOL1*-related alleles linked to kidney disease (offering protection from *Trypanosoma brucei gambiense* infection), and *HLA-B**5701 (associated with

hypersensitivity to the antiretroviral drug abacavir), which also showed increased frequency in region-specific African populations [8].

The evolutionary history of high-altitude and transitional populations shows a complex interaction of migration, adaptation, introgression, and disease-related genetic variation. Convergence evolution contributed to the adaptation of the highland regions and was a hot research topic in human genetics [25,40,47,68]. In mainland Southeast Asia, genetic diversity reflects mixing between ancient lineages and recent migrations, with adaptive genes linked to diet, climate, and pathogen exposure [66]. Past selective pressures continue to influence current risks for metabolic, immune, and cardiovascular disorders in ethnolinguistically diverse groups, including our reported TYC populations [12]. Comparative studies of high-altitude adaptation in the Andes [69], Papua New Guinea [70], and Tibet [25,40,47,68] have identified convergent selection on hypoxia-related pathways, although through different genetic mechanisms shaped by drift, gene flow, and local environmental factors. In the Andes, an *EPAS1*/*HIF2A* missense variant affects hematocrit, balancing oxygen delivery and blood viscosity [69]. Tibetans carry unique adaptive alleles, including *EPAS1* [40] and MCUR1 [71], that regulate hypoxia response, red blood cell production, and cardiovascular function. Himalayan groups like the Deng show altitude-related adaptations through different gene-driven mechanisms, further illustrating the variety of adaptive solutions [25]. Pharmacogenomic profiling has also uncovered strong gene–drug interactions in TYC populations. Warfarin dose needs vary significantly among groups, with EA individuals generally requiring lower doses to minimize bleeding risks. *VKORC1*-rs9923231-C-T is found at a higher frequency in eastern China than in western regions, highlighting detailed pharmacogenetic differences and the importance of including underrepresented populations in clinical dosing guidelines. Therefore, collecting WGS resources from underrepresented populations and conducting extensive sampling to identify local substructure are vital for advancing precision medicine. The genetic makeup of TYC populations likely includes a mosaic of ancient variants, local adaptations, and ongoing selective processes. Variants considered pathogenic might serve as adaptive trade-offs; alleles affecting adrenal steroid metabolism or peripheral nerve function could enhance energy efficiency, thermoregulation, or pathogen resistance in high-altitude environments [12]. These pleiotropic architectures emphasize the importance of context when assessing deleteriousness, influencing variant classification, ClinVar interpretation, and the reliability of genomic medicine across populations. Integrating evolutionary insights with medical genomics improves understanding of health risks and resilience in underserved groups, increasing the accuracy of variant interpretation in TYC populations. This approach enhances the translational value of genomic research, supporting precision medicine in China and around the world.

Limitations

Despite extensive sampling, WGS, and modeling, gaps remain. While 1,031 genomes have been produced, many ethnolinguistic TYC groups, especially small, isolated populations like the Mosuo, Dulong, and Jiarong, are underrepresented. These groups are crucial for capturing genetic diversity and need deeper sequencing and references like T2T and pangenomes [58]. Lack of Y-chromosome and mitochondrial data hampers understanding of these populations, sex-biased admixture, and adaptations [34,68]. Ancient DNA from TYC and neighboring areas is unevenly distributed; key Neolithic and Bronze Age genomes exist, but older sequences from core TYC sites are lacking, limiting insights into population history. High-quality ancient DNA from sites like Gaoshan and Haimenkou would improve models of admixture and divergence. Besides, many loci linked to adaptation and archaic introgression remain unconfirmed; current evidence relies on statistical signals and bioinformatics, needing functional validation like gene expression, CRISPR, and physiological studies, as in Tibetans' hypoxia research [39,71]. Limited reference genomes bias introgression analyses, possibly missing unsampled lineages. GWAS from non-TYC groups need validation in TYC datasets. Demographic models often oversimplify complex migrations; adding rare variants, ancient DNA, and continuous migration models could help. Pharmacogenomic and clinical research are limited due to scarce region-specific health data. Integrating health records, longitudinal studies, and local trials is crucial for personalized medicine in TYC communities.

Conclusion

We characterized the genomic diversity of indigenous populations across the TYC, yielding a high-resolution view of genetic makeup, evolutionary history, introgression, and adaptation. By analyzing genomes from 1,031 individuals, we uncovered millions of variants, including numerous novel and rare alleles, and resolved fine-scale geographic and linguistic substructure. We detect complex signatures of adaptation and archaic introgression, identify markers associated with high-altitude physiology, and trace substantial ancestry to Neolithic millet farmers. Together, these results delineate distinct genetic groupings shaped by ancient migrations and local environments and demonstrate the clinical and

translational value of WGS in diverse populations. Expanding high-depth WGS across broader geographic and ethnolinguistic ranges, and integrating T2T and pangenome reference genomes, will sharpen detection of complex variants and enhance inference of population history. Our findings underscore the need to include underrepresented groups in genomic research to refine disease-risk assessments, advance equitable genomic medicine, and deepen understanding of human evolution.

Materials and methods

Population samples and ethical statement

A total of 147 peripheral blood samples were collected from 126 individuals of Tibeto-Burman descent and 22 individuals of Austroasiatic descent. The Tibeto-Burman speakers included 44 Qiang, 16 Amdo Tibetan, and 22 BMT individuals from Jiuzhaigou County, Aba Tibetan and Qiang Autonomous Prefecture, Sichuan Province, China, along with 15 Bai, 12 Hani, 5 Lahu, and 12 Pumi individuals from Dali, Honghe, Puer, and Nujiang in Yunnan Province, respectively. The Austroasiatic-speaking individuals consisted of 11 Blang from Xishuangbanna and 10 Wa from Lincang, Yunnan Province, China. Additionally, saliva samples were collected from 884 Sinitic speakers from Shandong, Shanxi, and Henan Provinces through comprehensive geographic sampling. All participants were offspring of nonconsanguineous marriages of indigenous residents within three generations. All procedures adhered to the recommendations of the Declaration of Helsinki, which was revised in 2000 [72]. The study protocols were reviewed and approved by the Ministry of Science and Technology of the Human Genetic Resources Administration of China (HGRAC), registration number 2024SQCJ000261, and the Medical Ethics Committee of West China Hospital of Sichuan University, approval number 2024-173.

Genome sequencing and data processing

Genomic DNA was extracted from the peripheral blood of 147 Tibeto-Burman and Austroasiatic individuals via the QIAamp DNA Mini Kit (QIAGEN, Germany) according to established protocols. WGS was performed on the DNBSEQ-T7 platform (MGI, Shenzhen, China), yielding a mean depth of 12.5× for 150 bp paired-end reads, with standard library preparation. Each TYC sample was processed in a separate lane, generating at least 30 GB of data that passed the filtering process. Quality control ensured that 80% of the bases achieved a base quality score of at least 30. The reads were merged, adaptors were trimmed, and the sequences were aligned to the human reference genome via BWA v.0.7.13 [73]. Variant calling was conducted with the HaplotypeCaller module in the Genome Analysis Toolkit (GATK) [74]. For downstream analysis, only biallelic variants were retained, resulting in a final dataset of 26,141,086 SNPs. Additionally, 884 Han Chinese individuals, reported here for the first time, were genotyped via the Infinium Global Screening Array (Illumina, CA, USA) and aligned to the GRCh37 reference genome. Genetic relatedness between individuals was estimated via PLINK v.1.90 [75] and King [76]. SNPs with a missing call rate greater than 0.05 or those failing Hardy–Weinberg equilibrium testing (--geno 0.05 and --hwe 10-6) were removed, along with individuals with genotyping rates exceeding 0.05 (--mind 0.05).

Population dataset

i. WGS dataset: WGS data from phase 3 of the 1KGP [46] and HGDP [44] were used for genotype calling and quality control alongside newly sequenced TYC samples, which were mapped to the human reference genome GRCh38. This resulted in a WGS dataset comprising 964 samples and 26,141,086 variants, which revealed genetic variants, selective signatures, and medically relevant variations.

ii. 2240K dataset: PileupCaller v1.5.2 was used to call 2240K SNVs per sample from BAM files mapped to the human reference genome (GRCh37). To facilitate a comprehensive spatiotemporal analysis of population history, this dataset was merged with the publicly available 1240K dataset from the Allen Ancient DNA Resource (AADR, https://reich.hms.harvard.edu/allen-ancient-dna-resource-aadr-downloadable-genotypes-present-day-and-ancient-dna-data), resulting in the 2240K dataset, which contains 1,135,618 SNPs. Further integration with the publicly available Human Origin (HO) dataset from the David Reich Laboratory (https://reich.hms.harvard.edu/datasets) has generated the 2240K_HO dataset, which contains 593,050 SNPs.

iii. WGS (GRCh37) dataset: BAM files generated by mapping raw sequencing data to the human reference genome GRCh37 were utilized to simulate demographic history. Additionally, VCF files were used to identify archaic introgression, integrating data from Altai and Vindija Neanderthal [77], as well as Altai Denisovan [78] individuals.

iv. Illumina dataset: Previously reported populations genotyped via the Illumina chip from the 10K_CPGDP dataset were combined with newly studied Han Chinese individuals to form the Illumina dataset, which includes 516,448 SNPs. Further integration with publicly available datasets from the HGDP/Oceania genomic resource, the HO dataset, and the 2240K dataset generated the Illumina_HGDP,

Illumina_HO, and Illumina_2240K datasets, which contained 460,678, 55,649, and 146,802 SNPs, respectively.

Variant quality control and annotations

Following stringent quality control used in 10K_CPGDP [45], variants were categorized into SNPs and InDels via VCFtools (version 0.1.16; http://vcftools.sourceforge.net/) based on the following criteria: (1) sites with a quality score below 30 were excluded, and (2) sites with a P value less than 1e-10, indicating deviation from HWE, were also removed. After applying these filters, allele counts were calculated, and only biallelic SNPs were retained for further analysis, resulting in 9.8 million identified biallelic SNPs. The clinical significance of the identified SNPs and their GWAS hits were annotated via ClinVar (version 20240419; https://www.ncbi.nlm.nih.gov/clinvar/) and the GWAS catalog (version 2024-04-23; https://www.ebi.ac.uk/gwas/), respectively. The dbSNP dataset (build 156; https://ftp.ncbi.nih.gov/snp/archive/) was consulted to determine whether the identified SNPs in the 147 WGS samples had been previously reported. Variants were classified as known if they matched a dbSNP variant in terms of genomic position, reference allele, and alternative allele; otherwise, they were considered novel.

Population genetic analysis

Principal component analysis

PCA was performed via the smartPCA program from EIGENSOFT [79], with parameters set to numoutlieriter: 0 and lsqproject: YES. To investigate the genetic relationships between TYC individuals and global populations, a worldwide PCA was conducted using a WGS dataset (GRCh38) with biallelic autosomal common variants (MAF > 0.05). Additionally, a series of PCAs were carried out to examine population structure within the context of ancient and modern EA and mainland SEA Tibeto-Burman, Austroasiatic, and Sinitic speakers. In these analyses, the studied populations were iteratively refined by removing "outliers" based on the first two PCs via the Illumina_2240K dataset, which projected selected ancient individuals onto the background of modern populations.

Genetic difference analyses

Pairwise fixation indices (F_{ST}) were calculated via PLINK v.1.90 [75] to estimate the F_{ST}-based genetic distances between the studied TYC/Sinitic individuals and various geographic and linguistic reference populations. These analyses were based on the Illumina_2240K dataset, with all the samples included.

Model-based ADMIXTURE analysis

For admixture analysis, quality control was performed on the WGS dataset and the Illumina_2240K dataset using pruned SNPs in strong linkage disequilibrium. The parameters for pruning were set to --indep-pairwise 50 10 0.1, and analyses were conducted in PLINK v.1.90 [75]. ADMIXTURE v.1.3.0 was run 100 times with a random seed for each value of K from 2 to 20, generating cross-validation error estimates to determine the optimal value of K [80].

Testing for the presence of admixture

To investigate and validate genetic drift between the TYC people and reference populations, we calculated outgroup f_3-statistics via f_3(Reference, TYC; Mbuti) based on the qp3Pop module in ADMIXTOOLS [81]. Admixture f_3-statistics, in the form of f_3(Source1, Source2, TYC), were also conducted to explore potential admixture signals in the focal populations, with significantly negative f_3 values (Z scores less than -3) indicating evidence of admixture between two predefined ancestry surrogates. Additionally, qpDstat software in ADMIXTOOLS was employed to assess gene flow events via the f_4:YES parameters. Three general models were constructed for analysis. First, the symmetrical f_4(TYC1, TYC2; Reference, Mbuti) was used to examine whether two sublineages of the TYC population formed a clade relative to the reference populations. The asymmetrical f_4(Reference1, TYC; Reference2, Mbuti) and affinity f_4(Reference1, Reference2; Target, Mbuti) methods were applied to test whether the TYC population shares more alleles with Reference1 or Reference2 than other populations do.

Runs of homozygosity and nucleotide diversity estimation

To detect potential consanguinity within populations from the Illumina_2240K dataset, we employed PLINK v.1.9 [75] to identify ROH. The sliding window size was set to 500 kb, containing a minimum of 50 SNPs, allowing 1 heterozygote and up to 5 missing calls per window. The total ROH length was then calculated to enable a comparative analysis of distribution patterns across populations at the population level. Nucleotide diversity was computed via VCFtools [82], with estimates based on average pairwise differences within each population for every 20 kb window, with no overlap.

Demographic history modeling with qpAdm and qpWave

To assess ancestry proportions and admixture patterns among the thirteen populations inhabiting TYC, we conducted qpAdm analyses via ADMIXTOOLS2 [81]. These populations were modeled as mixtures of predefined source populations, where the ancestry of target populations was estimated using a set of

source populations (left populations) and a set of reference populations (right populations), without requiring an explicit phylogenetic relationship between them. The reference populations, selected for their distinct relationships with the target populations, included Mbuti, Papuan, Ust_Ishim, Onge, Mixe, Villabruna, GanjDareh, and Kostenki14. qpAdm models were accepted when the p value exceeded 0.05, provided that the admixture proportion and its standard error were positive; models with p values below 0.01 were rejected. Prior to qpAdm modeling, qpWave analyses were performed to ensure that the reference populations were sufficiently differentiated based on their unique genetic affinities and relationships with the outgroup. The outgroups, as identified in previous studies [22], included Mbuti, Afanasievo, Papuan, Jomon, China_AR_EN, Longlin, China_NEastAsia_Coastal_EN, China_SEastAsia_Island_EN, and Hoabinhian.

Admixture dating

To provide further insights into the historical context of admixture events, MALDER (v.1.0) [83] was utilized to estimate the timing of multiple admixture events within TYC populations. A minimum genetic distance of 0.005 cM was set for curve fitting to account for short-range linkage disequilibrium. Haplotype-sharing patterns were examined via ChromoPainter, which is based on the Illumina_HGDP dataset, to detect potential admixture within TYC populations. Subsequently, fastGLOBETROTTER [84] was applied by modeling the painting profile of a target cluster, as inferred by ChromoPainter, as a mixture of those from a set of reference clusters. The donor populations included Han, Yao, Li, Tibetan, Daur, Hezhen, Basque, French, Balochi, Brahui, Pathan, Sindhi, Yakut, and Cambodian. In fastGLOBETROTTER, 100 bootstrap replicates were performed to assess the statistical significance of the admixture event and to establish a 95% confidence interval for dating the admixture. We applied the formula $y = 1950 - 29*(g - 1)$, where y represents the year of admixture and g is the estimated number of generations.

Identification by descent and fine-STRUCTURE analyses

Haplotypes were estimated via SHAPEIT v.5.0, which is based on recommended human genetic maps [85]. Shared IBD segments within and between individuals or populations were detected via Refine IBD [86]. The coancestry matrix was calculated, and the fine-scale population structure was explored via fineSTRUCTURE v.4.0 with default parameters.

Estimation of Ne and divergence time

The historical effective population size (Ne) for each Tibeto-Burman/Austroasiatic group was inferred via the SMC++ program [87], with analyses performed on all available genomes. MSMC2 was then applied to infer the divergence time of the TYC population [88], with two samples randomly selected from each population. Initially, sample-specific VCF and mask files were generated via bamCaller.py from the MSMC-tools package. Multistep files for two-phased diploid individuals, including target TYC individuals and Han individuals from 1KGP, were then produced, and VCF and mask files were merged via the generate_multihetsep.py script. MSMC2 was subsequently run to estimate coalescence rates, and the combineCrossCoal.py repository was used to create a joint output file. The relative cross-coalescence rate (rCCR) was computed, and the divergence time was estimated when the rCCR reached 0.5. For all the demographic analyses, the results were scaled, assuming a mutation rate of 1.25×10^{-8} per base per generation and a generation time of 29 years. Both SMC++ and MSMC2 analyses were performed via the WGS (GRCh37) dataset.

Biological adaptation

Identification of positive natural selection

For natural selection analyses, allele frequency-based (PBS) and haplotype-based (iHS) methods were applied via the WGS dataset. The PBS approach was first employed to detect signals of natural selection in the TYC population, using the Chinese Han (CHB) population from Beijing as the ingroup and the WGS data of the French, Basque, Sardinian, Italian, and Tuscan populations from HGDP (EUR) as the outgroup reference. SNPs with extreme branch lengths (top 0.1%) were considered strong candidates for the genetic basis of adaptive evolution. Haplotype-based tests were then performed to confirm these selection signals, with iHS estimated via selscan v.1.2.078 for each lineage of the TYC population. To detect local adaptive signals, we normalized iHS scores across 20 kb windows, excluding windows with fewer than 20 SNVs. Based on the background distribution of normalized scores across all TYC lineages, windows with $|iHS| > 2$ were selected. Additionally, PBS was conducted on the Sinitic population from the Central Plain to detect selection signals via the Illumina_HGDP dataset, with the Miao and EUR populations serving as ingroup and outgroup references, respectively.

Functional annotation of natural selection signatures

Functional enrichment analysis of genes with strongly selective signals was conducted via Metascape (https://metascape.org/gp/index.html#/main/step1), an online tool that integrates widely used functional category ontologies. Previously reported genome-wide associations for genes and variants were

identified through the GWAS Catalog database.

Medically relevant variants and pharmacogenomic

The WGS dataset was utilized to catalog medically relevant variants. Medically actionable variants were identified by annotating SNVs based on the gene list specified by the ACMG for reporting secondary findings in clinical exome and genome sequencing [49]. The genetic burden associated with the ACMG gene panel version 3.2 was assessed via the following metrics: total number of observed alleles, median and range of alleles per individual, the proportion (%) of samples carrying at least one allele, and the genes in which variants were detected. Clinical annotation of SNVs was performed via the NCBI ClinVar database (https://www.ncbi.nlm.nih.gov/clinvar/), a publicly accessible resource linking medically relevant variants to phenotypes ranging from 0-Uncertain significance to 5-Pathogenic. The burden of these variants was also calculated, including the total allele count, median (range) number of alleles per sample, and number of samples with at least one allele and genes, which was consistent with the ACMG variant analysis. To explore the pharmacogenomic landscape, 23 CPIC drug–gene pairs were compiled in conjunction with PharmGKB clinical annotation levels 1A/1B, defined as follows: (Level 1A) gene–drug pairs with variant–specific prescribing guidance in clinical guidelines or FDA-approved drug labels, supported by at least one clinical annotation; (Level 1B) gene–drug pairs without variant–specific guidance but backed by high-level evidence from at least two independent studies affirming the association [89].

Detection of introgressed sequences and variants

To identify sequences likely introgressed from Neanderthals or Denisovans, Sprime [54] and IBDmix [90] were applied to the WGS (GRCh37) dataset. Combining multiple methods for detecting archaic introgressed sequences has been shown to improve the detection rate of authentic introgressed segments [91]. To minimize the false positive rate, only shared introgressed segments were retained to establish a high-confidence set of introgressed sequences. SPrime was used without an archaic reference genome to identify introgressed segments. Since SPrime cannot handle missing data, sites with missing genotypes were excluded. A threshold of 140,000 for the SPrime score was set based on score distribution across populations, with only introgressed haplotypes containing more than 30 sites considered, as longer haplotypes provide stronger evidence of archaic origin. Neanderthal-specific segments were defined by a match rate to Neanderthal of 0.6 or higher and a Denisovan match rate below 0.3. Conversely, Denisovan-specific segments were identified with a Denisovan match rate above 0.3 and a Neanderthal match rate below 0.3.

Additionally, IBDMix, a computational approach utilizing an archaic reference genome, was employed to detect introgressed archaic haplotypes within phased genomic data. Phasing was conducted via Beagle 2.0 with default parameters. Prior to the analysis, genomic regions inaccessible in Vindija and Altai Neanderthals, as well as Altai Denisovan, were masked. The following filters were applied: (1) Coverage filter stratified by GC content, (2) minimum coverage of 10, (3) Heng Li's Mapability 35 (map35_100), (4) mapping quality score threshold of 25, and (5) exclusion of tandem repeats, InDels, and sites located within CpG islands. For the introgressed sequences, a final set of segments was curated using a logarithm of odds threshold greater than 4 and a minimum segment length of 50 kilobases. Pathway analysis was performed to identify biological pathways enriched in genes containing archaic variants via Matascape (https://metascape.org/gp/index.html#/main/step1), and associations with archaic variants were confirmed via GWAS summary statistics (https://www.ebi.ac.uk/gwas/).

Declarations

Ethnic approval and consent to participate

The protocols were reviewed and approved by the Ministry of Science and Technology of the Human Genetic Resources Administration of China (HGRAC; registration number 2024SLCJ000164) and the Medical Ethics Committee of West China Hospital of Sichuan University (approval number 2024-173). The study was conducted in accordance with the principles of the Helsinki Declaration.

Availability of data and materials

All the data generated or analyzed during this study are included in this published article, its supplementary information files, and publicly available repositories.

The genome-wide variation data were obtained from the public dataset of the Allen Ancient DNA Resource (AADR) in the Reich Laboratory (https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/FFIDCW). The results of the analyses have been submitted to the supplementary materials and deposited into the OMIX database (https://ngdc.cncb.ac.cn/omix/) under accession number OMIX007569. The whole-genome sequencing

data from the 147 Tibeto-Burman and Austroasiatic individuals sampled in the Tibetan-Yi Corridor, generated in this study, have been deposited in the Genome Variation Map [92] in the National Genomics Data Center, Beijing Institute of Genomics, Chinese Academy of Sciences and China National Center for Bioinformation [93] under accession number GVM000862 (https://ngdc.cncb.ac.cn/gvm/getProjectDetail?project=GVM000862). The variation data of 884 Sinitic speakers reported in this paper have been deposited in the Genome Variant Map under accession number GVM000886 (https://ngdc.cncb.ac.cn/gvm/getProjectDetail?project=GVM000886). Data access and use are restricted to academic research in population genetics, including research on population origins, ancestry and history. These data are available under restricted access due to the sensitive information they contain, which could compromise the privacy and consent of the research participants. Requests for access to data can be directed to Guanglin He (Guanglinhescu@163.com). The access to and use of data must comply with the regulations of the People's Republic of China on the administration of human genetic resources (No. 2025BAT00381 and 2025BAT00386).

Credit authorship contribution statement

Chao Liu, Renkuan Tang, Mengge Wang, **Libing Yun**, **and Guanglin He**: Conceptualization, project administration, supervision, funding acquisition. **Mengge Wang**, **Shuhan Duan, Yunhui Liu, Lintao Luo, and Guanglin He**: writing - original draft, review & editing Methodology, Data analysis, Investigation, Sampling, Review & Editing. **Shuhan Duan, Yunhui Liu, Lintao Luo, Qiuxia Sun and Yuntao Sun**: Writing - Review & editing. **Mengge Wang**, **Libing Yun, Qiuxia Sun, and Yuntao Sun:** Collected the samples, genomic DNA extraction and sequencing. **Chao Liu, Renkuan Tang, Mengge Wang**, **Libing Yun**, **Shuhan Duan, Yunhui Liu, and Guanglin He**: Supervision, review and revision of the manuscript.

Conflict of interest

The authors claim that no conflicts of interest exist.

Acknowledgments

This study was supported by the National Natural Science Foundation of China (82402203 and 82202078), the Major Project of the National Social Science Foundation of China (23&ZD203), the Open Project of the Key Laboratory of Forensic Genetics of the Ministry of Public Security (2022FGKFKT05), the Center for Archaeological Science of Sichuan University (23SASA01), the 1.3.5 Project for Disciplines of Excellence, West China Hospital, Sichuan University (ZYJC20002), and the Sichuan Science and Technology Program (2024NSFSC1518).

References

1 Sagart, L. *et al.* Dated language phylogenies shed light on the ancestry of Sino-Tibetan. *Proc Natl Acad Sci U S A* **116**, 10317-10322 (2019). https://doi.org/10.1073/pnas.1817972116

2 Zhang, M., Yan, S., Pan, W. & Jin, L. Phylogenetic evidence for Sino-Tibetan origin in northern China in the Late Neolithic. *Nature* **569**, 112-115 (2019). https://doi.org/10.1038/s41586-019-1153-z

3 Zhang, Z. *et al.* The Tibetan-Yi region is both a corridor and a barrier for human gene flow. *Cell Rep* **39**, 110720 (2022). https://doi.org/10.1016/j.celrep.2022.110720

4 Leipe, C., Long, T., Sergusheva, E. A., Wagner, M. & Tarasov, P. E. Discontinuous spread of millet agriculture in eastern Asia and prehistoric population dynamics. *Science advances* **5**, eaax6225 (2019). https://doi.org/10.1126/sciadv.aax6225

5 Fuller, D. Q. Pathways to Asian Civilizations: Tracing the Origins and Spread of Rice and Rice Cultures. *Rice* **4**, 78-92 (2012). https://doi.org/10.1007/s12284-011-9078-7

6 He, G. *et al.* Fine-scale genetic structure of Tujia and central Han Chinese revealing massive genetic admixture under language borrowing. *Journal of Systematics and Evolution* **59**, 1-20 (2021).

7 Consortium, U. K. B. W.-G. S. Whole-genome sequencing of 490,640 UK Biobank participants. *Nature* (2025). https://doi.org/10.1038/s41586-025-09272-9

8 Choudhury, A. *et al.* High-depth African genomes inform human migration and health. *Nature* **586**, 741-748 (2020). https://doi.org/10.1038/s41586-020-2859-7

9 Silcocks, M. *et al.* Indigenous Australian genomes show deep structure and rich novel variation. *Nature* **624**, 593-601 (2023). https://doi.org/10.1038/s41586-023-06831-w

10 Verma, A. *et al.* Diversity and scale: Genetic architecture of 2068 traits in the VA Million Veteran Program. *Science* **385**, eadj1182 (2024). https://doi.org/10.1126/science.adj1182

11 All of Us Research Program Genomics, I. Genomic data in the All of Us Research Program.

Nature **627**, 340-346 (2024). https://doi.org/10.1038/s41586-023-06957-x
12 Benton, M. L. *et al.* The influence of evolutionary history on human health and disease. *Nat Rev Genet* **22**, 269-283 (2021). https://doi.org/10.1038/s41576-020-00305-9
13 Yang, Q. *et al.* The large-scale whole-genome sequencing era expedited medical discovery and clinical translation. *EngMedicine* **2**, 100055 (2025). https://doi.org/10.1016/j.engmed.2025.100055
14 Cao, Y. *et al.* The ChinaMAP analytics of deep whole genome sequences in 10,588 individuals. *Cell Res* **30**, 717-731 (2020). https://doi.org/10.1038/s41422-020-0322-9
15 Zhang, P. *et al.* NyuWa Genome resource: A deep whole-genome sequencing-based variation profile and reference panel for the Chinese population. *Cell Rep* **37**, 110017 (2021). https://doi.org/10.1016/j.celrep.2021.110017
16 Yang, M. Y. *et al.* SEAD reference panel with 22,134 haplotypes boosts rare variant imputation and genome-wide association analysis in Asian populations. *Nat Commun* **15**, 10839 (2024). https://doi.org/10.1038/s41467-024-55147-4
17 Zhou, D. *et al.* Non-coding genetic elements of lung cancer identified using whole genome sequencing in 13,722 Chinese. *Nat Commun* **16**, 7365 (2025). https://doi.org/10.1038/s41467-025-62459-6
18 Wang, Z. *et al.* YanHuang Paternal Genomic Resource Suggested A Weakly-Differentiated Multi-Source Admixture Model for the Formation of Han's Founding Ancestral Lineages. *Genomics, Proteomics & Bioinformatics* (2025). https://doi.org/10.1093/gpbjnl/qzaf049
19 Wang, M. *et al.* Genome-driven Chinese precision medicine: Biobank-scale genomic research as a new paradigm. *The Innovation Life* **3** (2025). https://doi.org/10.59717/j.xinn-life.2025.100131
20 Luo, L. *et al.* Sequencing and characterizing human mitochondrial genomes in the biobank-based genomic research paradigm. *Sci China Life Sci* **68**, 1610-1625 (2025). https://doi.org/10.1007/s11427-024-2736-7
21 Wang, T. *et al.* Prehistoric genomes from Yunnan reveal ancestry related to Tibetans and Austroasiatic speakers. *Science* **388**, eadq9792 (2025). https://doi.org/10.1126/science.adq9792
22 Tao, L. *et al.* Ancient genomes reveal millet farming-related demic diffusion from the Yellow River into southwest China. *Curr Biol* **33**, 4995-5002 e4997 (2023). https://doi.org/10.1016/j.cub.2023.09.055
23 Zhang, C. *et al.* Differentiated demographic histories and local adaptations between Sherpas and Tibetans. *Genome Biol* **18**, 115 (2017). https://doi.org/10.1186/s13059-017-1242-y
24 Zheng, W. *et al.* Large-scale genome sequencing redefines the genetic footprints of high-altitude adaptation in Tibetans. *Genome Biol* **24**, 73 (2023). https://doi.org/10.1186/s13059-023-02912-1
25 Ge, X. *et al.* Genetic Origins and Adaptive Evolution of the Deng People on the Tibetan Plateau. *Mol Biol Evol* **40** (2023). https://doi.org/10.1093/molbev/msad205
26 Kamm, J., Terhorst, J., Durbin, R. & Song, Y. S. Efficiently inferring the demographic history of many populations with allele count data. *J Am Stat Assoc* **115**, 1472-1487 (2020). https://doi.org/10.1080/01621459.2019.1635482
27 Quintana-Murci, L. Genetic, Linguistic and Archaeological Perspectives on Human Diversity in Southeast Asia. *The American Journal of Human Genetics* **71**, 1253-1255 (2002). https://doi.org/10.1086/343817
28 Yang, Z. *et al.* Genetic evidence of tri-genealogy hypothesis on the origin of ethnic minorities in Yunnan. *BMC Biol* **20**, 166 (2022). https://doi.org/10.1186/s12915-022-01367-3
29 Lipson, M. *et al.* Ancient genomes document multiple waves of migration in Southeast Asian prehistory. *Science* **361**, 92-95 (2018). https://doi.org/10.1126/science.aat3188
30 McColl, H. *et al.* The prehistoric peopling of Southeast Asia. *Science* **361**, 88-92 (2018). https://doi.org/10.1126/science.aat3628
31 Sun, Q. *et al.* Differentiated adaptative genetic architecture and language-related demographical history in South China inferred from 619 genomes from 56 populations. *BMC Biol* **22**, 55 (2024). https://doi.org/10.1186/s12915-024-01854-9
32 He, G. *et al.* Differentiated genomic footprints suggest isolation and long-distance migration of Hmong-Mien populations. *BMC Biol* **22**, 18 (2024). https://doi.org/10.1186/s12915-024-01828-x
33 Singh, P. P. *et al.* Dissecting the paternal founders of Mundari (Austroasiatic) speakers associated with the language dispersal in South Asia. *Eur J Hum Genet* **29**, 528-532 (2021). https://doi.org/10.1038/s41431-020-00745-1
34 Wang, M. *et al.* Multiple Human Population Movements and Cultural Dispersal Events Shaped

the Landscape of Chinese Paternal Heritage. *Mol Biol Evol* **41** (2024). https://doi.org/10.1093/molbev/msae122
35 Sun, Y. *et al.* Distinguished biological adaptation architecture aggravated population differentiation of Tibeto-Burman-speaking people. *J Genet Genomics* **51**, 517-530 (2024). https://doi.org/10.1016/j.jgg.2023.10.002
36 Wang, M. *et al.* Forensic investigative genetic genealogy: expanding pedigree tracing and genetic inquiry in the genomic era. *J Genet Genomics* **52**, 460-472 (2025). https://doi.org/10.1016/j.jgg.2024.06.016
37 He, G. *et al.* Forensic characteristics and phylogenetic analyses of the Chinese Yi population via 19 X-chromosomal STR loci. *Int J Legal Med* **131**, 1243-1246 (2017). https://doi.org/10.1007/s00414-017-1563-0
38 He, G. *et al.* Population genomics of Central Asian peoples unveil ancient Trans-Eurasian genetic admixture and cultural exchanges. *hLife* **2**, 554-562 (2024). https://doi.org/10.1016/j.hlife.2024.06.006
39 Yang, Z. *et al.* Genetic adaptation of skin pigmentation in highland Tibetans. *Proc Natl Acad Sci U S A* **119**, e2200421119 (2022). https://doi.org/10.1073/pnas.2200421119
40 Simonson, T. S. *et al.* Genetic evidence for high-altitude adaptation in Tibet. *Science* **329**, 72-75 (2010). https://doi.org/10.1126/science.1189406
41 Huerta-Sanchez, E. *et al.* Altitude adaptation in Tibetans caused by introgression of Denisovan-like DNA. *Nature* **512**, 194-197 (2014). https://doi.org/10.1038/nature13408
42 Zhang, X. *et al.* The history and evolution of the Denisovan-EPAS1 haplotype in Tibetans. *Proc Natl Acad Sci U S A* **118** (2021). https://doi.org/10.1073/pnas.2020803118
43 Wu, X. & Yan, Y. The internal anatomical structure of the Ziyang human skull fossils. *Acta Anthropologica Sinica* **39**, 10 (2020).
44 Bergstrom, A. *et al.* Insights into human genetic variation and population history from 929 diverse genomes. *Science* **367** (2020). https://doi.org/10.1126/science.aay5012
45 He, G. *et al.* Pilot work of the 10K Chinese People Genomic Diversity Project along the Silk Road suggests a complex east-west admixture landscape and biological adaptations. *Sci China Life Sci* **68**, 914-933 (2025). https://doi.org/10.1007/s11427-024-2748-4
46 Genomes Project, C. *et al.* A global reference for human genetic variation. *Nature* **526**, 68-74 (2015). https://doi.org/10.1038/nature15393
47 Yi, X. *et al.* Sequencing of 50 human exomes reveals adaptation to high altitude. *Science* **329**, 75-78 (2010). https://doi.org/10.1126/science.1190371
48 Jonnalagadda, M. *et al.* MC1R diversity and its role in skin pigmentation variation in West Maharashtra, India. *American journal of human biology : the official journal of the Human Biology Council* **34**, e23734 (2022). https://doi.org/10.1002/ajhb.23734
49 Miller, D. T. *et al.* ACMG SF v3.2 list for reporting of secondary findings in clinical exome and genome sequencing: A policy statement of the American College of Medical Genetics and Genomics (ACMG). *Genetics in medicine : official journal of the American College of Medical Genetics* **25**, 100866 (2023). https://doi.org/10.1016/j.gim.2023.100866
50 Gao, Y. *et al.* PGG.Han: the Han Chinese genome database and analysis platform. *Nucleic Acids Res* **48**, D971-D976 (2020). https://doi.org/10.1093/nar/gkz829
51 GenomeAsia, K. C. The GenomeAsia 100K Project enables genetic discoveries across Asia. *Nature* **576**, 106-111 (2019). https://doi.org/10.1038/s41586-019-1793-z
52 Johnson, J. A. *et al.* Clinical Pharmacogenetics Implementation Consortium (CPIC) Guideline for Pharmacogenetics-Guided Warfarin Dosing: 2017 Update. *Clin Pharmacol Ther* **102**, 397-404 (2017). https://doi.org/10.1002/cpt.668
53 Qin, P. & Stoneking, M. Denisovan Ancestry in East Eurasian and Native American Populations. *Mol Biol Evol* **32**, 2665-2674 (2015). https://doi.org/10.1093/molbev/msv141
54 Browning, S. R., Browning, B. L., Zhou, Y., Tucci, S. & Akey, J. M. Analysis of Human Sequence Data Reveals Two Pulses of Archaic Denisovan Admixture. *Cell* **173**, 53-61 e59 (2018). https://doi.org/10.1016/j.cell.2018.02.031
55 Zhang, X. *et al.* The history and evolution of the Denisovan-EPAS1 haplotype in Tibetans. *Proc Natl Acad Sci U S A* **118**, e2020803118 (2021). https://doi.org/10.1073/pnas.2020803118
56 Wang, H. *et al.* Exposure to hypoxia causes stress erythropoiesis and downregulates immune response genes in spleen of mice. *BMC Genomics* **22**, 413 (2021). https://doi.org/10.1186/s12864-021-07731-x
57 Yang, Q. *et al.* High-quality Population-specific Haplotype-resolved Reference Panel in the Genomic and Pangenomic Eras. *Genomics, Proteomics & Bioinformatics* (2025). https://doi.org/10.1093/gpbjnl/qzaf022

58 Wang, M. *et al.* Mighty oaks from little acorns: High-quality genomes of underrepresented populations enhance health equity in precision medicine. *hLife* (2025). https://doi.org/10.1016/j.hlife.2025.05.014
59 Howley, C. *et al.* The expanding global genomics landscape: Converging priorities from national genomics programs. *Am J Hum Genet* **112**, 751-763 (2025). https://doi.org/10.1016/j.ajhg.2025.02.008
60 Wang, C. C. *et al.* Genomic insights into the formation of human populations in East Asia. *Nature* **591**, 413-419 (2021). https://doi.org/10.1038/s41586-021-03336-2
61 Liu, L., Chen, J., Wang, J., Zhao, Y. & Chen, X. Archaeological evidence for initial migration of Neolithic Proto Sino-Tibetan speakers from Yellow River valley to Tibetan Plateau. *Proc Natl Acad Sci U S A* **119**, e2212006119 (2022). https://doi.org/10.1073/pnas.2212006119
62 Shi, S. The Way of Multi − Ethnic Coexistence: Experience and Wisdom of Ethnic Exchanges in the Tibetan-Yi Corridor Region. *Journal of the Minzu University of China (Philosophy and Social Sciences Edition)* **50**, 57-64 (2023). https://doi.org/10.15970/j.cnki.1005-8575.2023.02.011
63 Hirota, K. & Semenza, G. L. Rac1 activity is required for the activation of hypoxia-inducible factor 1. *The Journal of biological chemistry* **276**, 21166-21172 (2001). https://doi.org/10.1074/jbc.M100677200
64 Choin, J. *et al.* Genomic insights into population history and biological adaptation in Oceania. *Nature* **592**, 583-589 (2021). https://doi.org/10.1038/s41586-021-03236-5
65 Karczewski, K. J. *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434-443 (2020). https://doi.org/10.1038/s41586-020-2308-7
66 He, Y. *et al.* Genome diversity and signatures of natural selection in mainland Southeast Asia. *Nature* **643**, 417-426 (2025). https://doi.org/10.1038/s41586-025-08998-w
67 Lin, X., Liu, Y., Wang, M., He, G. & Ma, Y. Genomic resources from the Yunnan-Guizhou Plateau reveal the linguistic-linked demographic history and insights into human adaptation and disease of multi-ancestry populations. *Journal of Systematics and Evolution* (2025). https://doi.org/10.1111/jse.70009
68 Peng, M. S. & Zhang, Y. P. Sex-biased adaptation shapes uniparental gene pools in Tibetans. *Sci China Life Sci* **67**, 611-613 (2024). https://doi.org/10.1007/s11427-023-2506-0
69 Lawrence, E. S. *et al.* Functional EPAS1/HIF2A missense variant is associated with hematocrit in Andean highlanders. *Science advances* **10**, eadj5661 (2024). https://doi.org/10.1126/sciadv.adj5661
70 Andre, M. *et al.* Positive selection in the genomes of two Papua New Guinean populations at distinct altitude levels. *Nat Commun* **15**, 3352 (2024). https://doi.org/10.1038/s41467-024-47735-1
71 Ping, J. *et al.* A highland-adaptation variant near MCUR1 reduces its transcription and attenuates erythrogenesis in Tibetans. *Cell Genom* **5**, 100782 (2025). https://doi.org/10.1016/j.xgen.2025.100782
72 World Medical Association, I. World Medical Association Declaration of Helsinki. Ethical principles for medical research involving human subjects. *Bull World Health Organ 2001* (2001).
73 Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**, 589-595 (2010). https://doi.org/10.1093/bioinformatics/btp698
74 DePristo, M. A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* **43**, 491-498 (2011). https://doi.org/10.1038/ng.806
75 Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**, 7 (2015). https://doi.org/10.1186/s13742-015-0047-8
76 Manichaikul, A. *et al.* Robust relationship inference in genome-wide association studies. *Bioinformatics* **26**, 2867-2873 (2010). https://doi.org/10.1093/bioinformatics/btq559
77 Prufer, K. *et al.* The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature* **505**, 43-49 (2014). https://doi.org/10.1038/nature12886
78 Meyer, M. *et al.* A high-coverage genome sequence from an archaic Denisovan individual. *Science* **338**, 222-226 (2012). https://doi.org/10.1126/science.1224344
79 Patterson, N., Price, A. L. & Reich, D. Population structure and eigenanalysis. *PLoS Genet* **2**, e190 (2006). https://doi.org/10.1371/journal.pgen.0020190
80 Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res* **19**, 1655-1664 (2009). https://doi.org/10.1101/gr.094052.109
81 Patterson, N. *et al.* Ancient admixture in human history. *Genetics* **192**, 1065-1093 (2012). https://doi.org/10.1534/genetics.112.145037

82 Danecek, P. *et al.* The variant call format and VCFtools. *Bioinformatics* **27**, 2156-2158 (2011). https://doi.org/10.1093/bioinformatics/btr330
83 Loh, P. R. *et al.* Inferring admixture histories of human populations using linkage disequilibrium. *Genetics* **193**, 1233-1254 (2013). https://doi.org/10.1534/genetics.112.147330
84 Wangkumhang, P., Greenfield, M. & Hellenthal, G. An efficient method to identify, date, and describe admixture events using haplotype information. *Genome Res* **32**, 1553-1564 (2022). https://doi.org/10.1101/gr.275994.121
85 Hofmeister, R. J., Ribeiro, D. M., Rubinacci, S. & Delaneau, O. Accurate rare variant phasing of whole-genome and whole-exome sequencing data in the UK Biobank. *Nat Genet* **55**, 1243-1249 (2023). https://doi.org/10.1038/s41588-023-01415-w
86 Browning, B. L. & Browning, S. R. Improving the accuracy and efficiency of identity-by-descent detection in population data. *Genetics* **194**, 459-471 (2013). https://doi.org/10.1534/genetics.113.150029
87 Terhorst, J., Kamm, J. A. & Song, Y. S. Robust and scalable inference of population history from hundreds of unphased whole genomes. *Nat Genet* **49**, 303-309 (2017). https://doi.org/10.1038/ng.3748
88 Schiffels, S. & Wang, K. MSMC and MSMC2: The Multiple Sequentially Markovian Coalescent. *Methods in molecular biology (Clifton, N.J.)* **2090**, 147-166 (2020). https://doi.org/10.1007/978-1-0716-0199-0_7
89 Whirl-Carrillo, M. *et al.* An Evidence-Based Framework for Evaluating Pharmacogenomics Knowledge for Personalized Medicine. *Clin Pharmacol Ther* **110**, 563-572 (2021). https://doi.org/10.1002/cpt.2350
90 Chen, L., Wolf, A. B., Fu, W., Li, L. & Akey, J. M. Identifying and Interpreting Apparent Neanderthal Ancestry in African Individuals. *Cell* **180**, 677-687 e616 (2020). https://doi.org/10.1016/j.cell.2020.01.012
91 Jacobs, G. S. *et al.* Multiple Deeply Divergent Denisovan Ancestries in Papuans. *Cell* **177**, 1010-1021 e1032 (2019). https://doi.org/10.1016/j.cell.2019.02.035
92 Li, C. *et al.* Genome Variation Map: a worldwide collection of genome variations across multiple species. *Nucleic Acids Res* **49**, D1186-D1191 (2021). https://doi.org/10.1093/nar/gkaa1005
93 Members, C.-N. Database resources of the national genomics data center, China National Center for Bioinformation in 2022. *Nucleic Acids Research* **50**, D27 (2022).

Figure legends

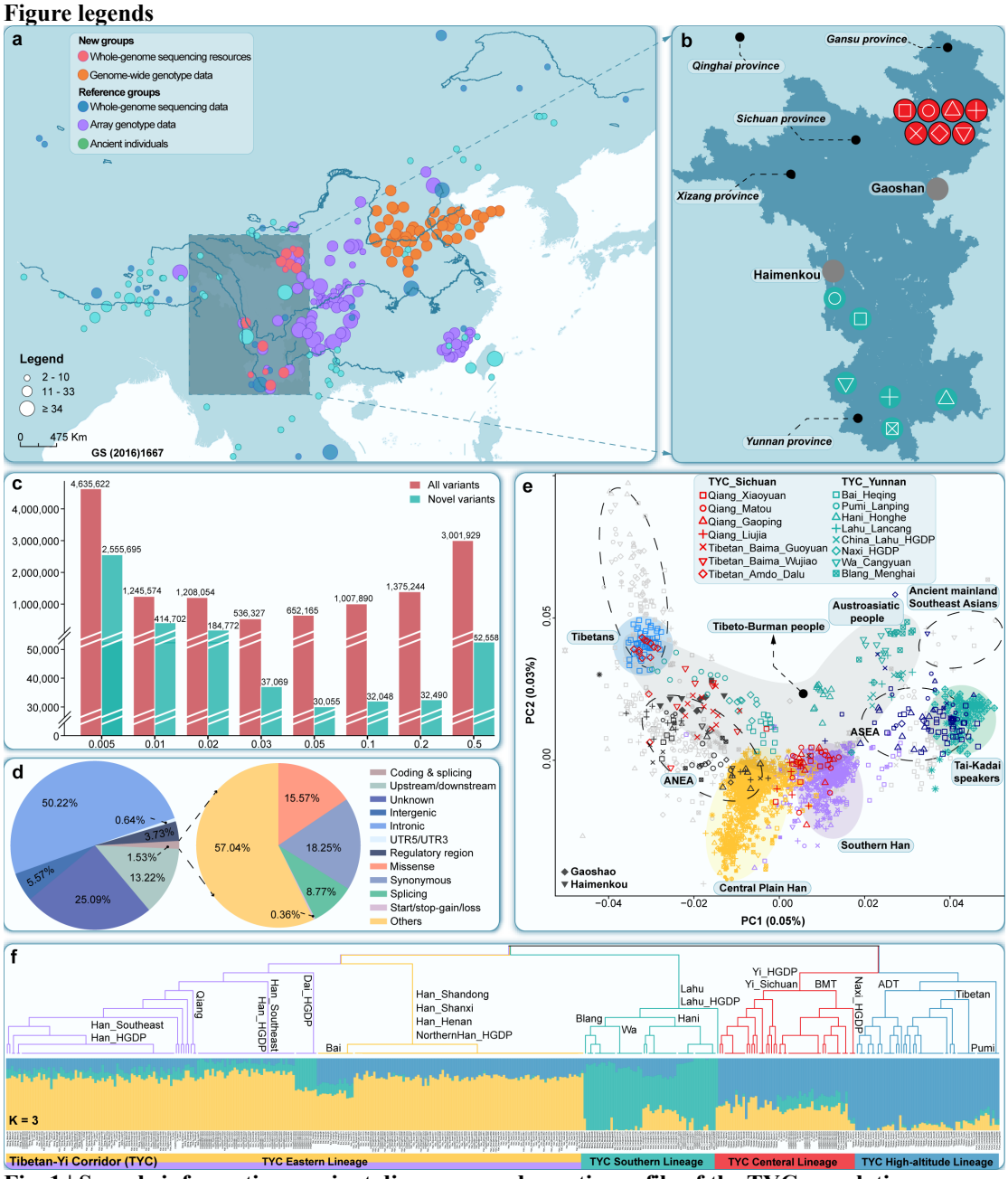

Fig. 1 | Sample information, variant discovery, and genetic profile of the TYC population.
(**a**) Geographical distribution of newly sampled Tibeto-Burman, Austroasiatic, and Sinitic populations, along with currently available genomic resources in East Asia (EA). (**b**) Geographic location of the Tibetan-Yi Corridor (TYC, dark blue color) and sampling populations involved in the whole-genome sequencing study. The symbols representing each population on the map correspond to the legend of the principal component analysis (PCA) shown in **Fig. 1e**, with more detailed individual information provided in **Table S1**. The base map was officially approved with the number GS (2016)1667 from the website http://bzdt.ch.mnr.gov.cn/. (**c**) The number of known (red) and novel (green) variants detected in the TYC populations, categorized based on dbSNP version 146, with the distribution classified by minor allele frequency. (**d**) Functional annotation of 13,662,805 biallelic SNPs, including 3.3 million previously unreported SNPs. (**e**) A two-dimensional principal component (PC) plot projects modern and ancient EA populations, as well as Southeast Asians (SEAs), within the EA context, highlighting clustering patterns among geographically and linguistically diverse EA populations. The modern reference EA populations, ancient Qinghai–Tibet Plateau individuals, and ancient mainland SEAs are colored gray. The detailed sample legends are shown in **Fig. S2e**. (**f**) A dendrogram constructed via haplotype-based fineSTRUCTURE and model-based ADMIXTURE results reveals clustering patterns among 23 EA populations based on the Illumina_2240K dataset. The ADMIXTURE results are displayed below, with each column representing an individual. The blue component is maximized in Amdo Tibetan (ADT), the

green component is enriched in Lahu, and the yellow component is prominent in the Central Plain Han (CPH). ANEA/ASEA, Ancient Northern/Southern East Asians.

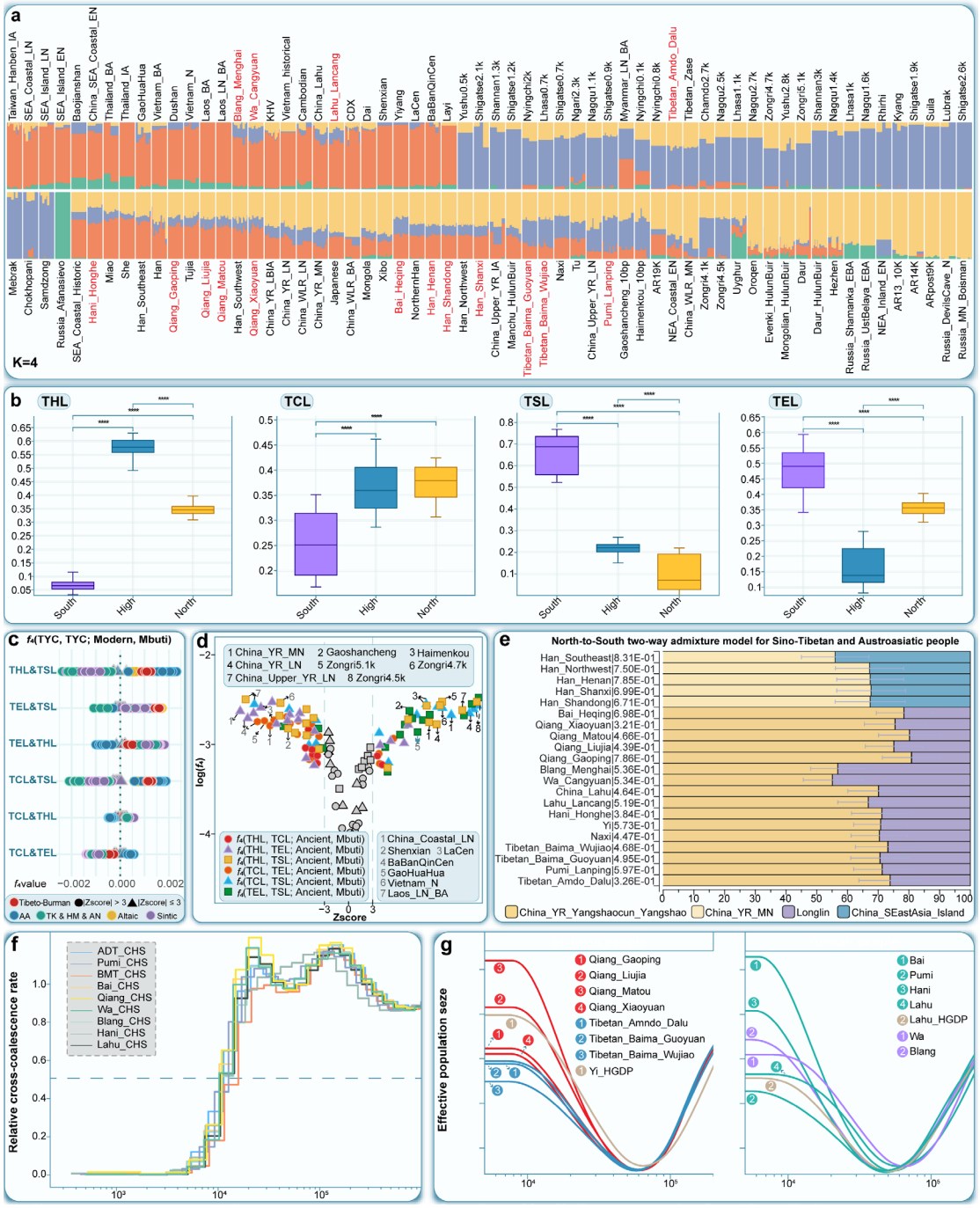

Fig. 2 | Genetic admixture and demographic history of TYC populations.
(**a**) Individual admixture proportions for TYC populations, Tibetans, present-day EA/SEA populations, ANEA, ASEA, and ancient SEA were estimated via the Illumina_2240K dataset. The minimum CV error for the four ancestral populations was selected, revealing northern ancestral components, including DevilsCave (yellow) and Shigatse 2.6K (blue), as well as southern ancestral components associated with Yiyang (red). (**b**) A Wilcoxon test was performed to assess the contributions of distinct ancestry components. The upper and lower borders represent the 95% confidence intervals. Statistical significance is denoted as follows: * for $P \leq 0.05$, ** for $P \leq 0.01$, *** for $P \leq 0.001$, and **** for $P \leq 0.0001$. (**c**) f_4-statistics, in the form of f_4(TYC1, TYC2; modern reference populations, Mbuti), are presented. Circles represent |Z-values| > 3, whereas triangles signify |Z-values| < 3. The colors correspond to classifications of ethnolinguistically diverse groups. (**d**) f_4-statistics for f_4(TYC1, TYC2; ancient reference populations, Mbuti) are displayed. The X-axis represents the logarithm of the f_4 values, with distinct shapes corresponding to different groups. The colored shapes indicate |Z-values| > 3, whereas the gray shapes

denote |Z-values| < 3. (**e**) Two-way admixture models of Tibeto-Burman and Austroasiatic speakers in TYC and Sinitic populations across geographically diverse regions were fitted via qpAdm, with well-fitting models illustrated. (**f**) MSMC2 cross-population results for pairs of TYC populations and Han Chinese (CHS) individuals were generated, assuming 25 years per generation and a mutation rate of 1.25 × 10^{-8}. (**g**) Effective population sizes for TYC populations were inferred via SMC++ and computed via composite likelihoods across all individuals per population. The curves for Qiangs and ADT from Sichuan are shown in red and blue, respectively, whereas Tibeto-Burman and Austroasiatic speakers from Yunnan are represented in green and purple, respectively. The HGDP populations are highlighted with a brown curve. Abbreviations: SEA, Southeast Asia; NEA, Northeast Asia; YR, Yellow River; IA, Iron Age; BA, Bronze Age; N, Neolithic; EN, Early Neolithic; MN, Middle Neolithic; LN, Late Neolithic; LBIA, Late Bronze or Iron Age; RCCR, Relative cross-coalescence rate; TB, Tibeto-Burman; AA, Austroasiatic; TK & HM & AN, Tai-Kadai, Hmong-Mien, and Austronesian.

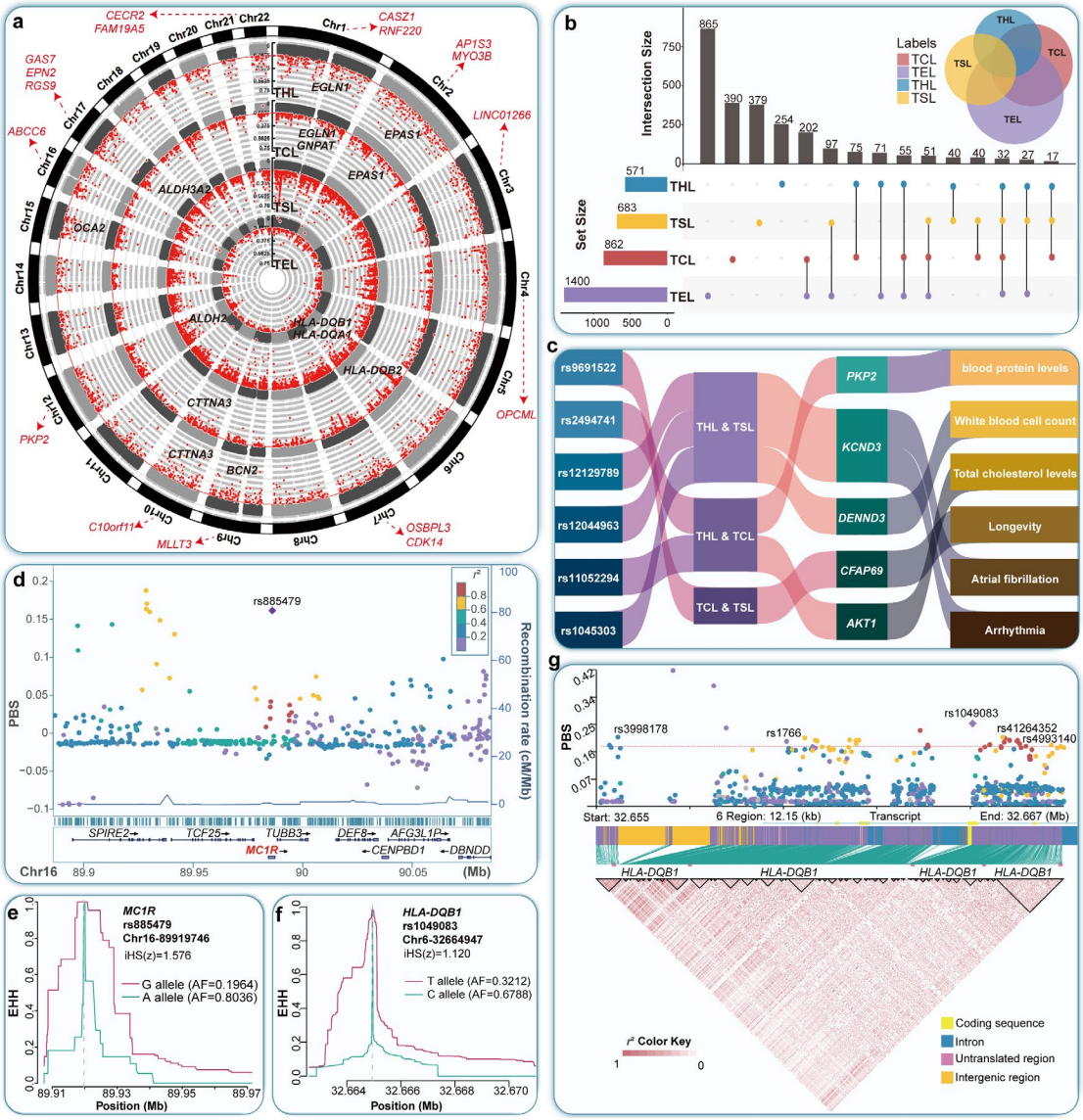

Fig. 3 | Selection signals identified in diverse genetic backgrounds of TYC individuals.
(**a**) Circular Manhattan plot depicting the population branch statistics (PBS) score distributions across various genetic backgrounds of TYC individuals. Loci within the top 0.1% of the PBS values are highlighted as red dots. Genes marked in red represent shared signatures with outlier scores in THL, TCL, and TSL, whereas those in black indicate population-specific selection signals. (**b**) Upset and Venn diagrams generated from the PBS results, with THL, TCL, TSL, and TEL colored blue, red, yellow, and purple, respectively. (**c**) A Sankey diagram illustrates the cluster projection of selective signals from genetically distinct TYC populations correlated with traits annotated via the GWAS Catalog. (**d**) The MC1R genomic region demonstrates strong linkage disequilibrium (LD) between rs885479, which is under positive selection, and adjacent loci. (**e**) Extended haplotype homozygosity (EHH) surrounding

MC1R-rs885479 indicates significant positive selection for the G allele. (**f**) Haplotype-based EHH analysis reveals a strong selection sweep acting on the T allele of *HLA-DQB1*-rs1049083. (**g**) Strong LD is observed between *HLA-DQB1*-rs1049083 and four previously reported SNPs associated with immunometabolism in TSL patients. Pairwise LD is measured as r^2 between SNPs, with red squares indicating the strength of the LD (dark red represents the strongest LD, and white represents the weakest).

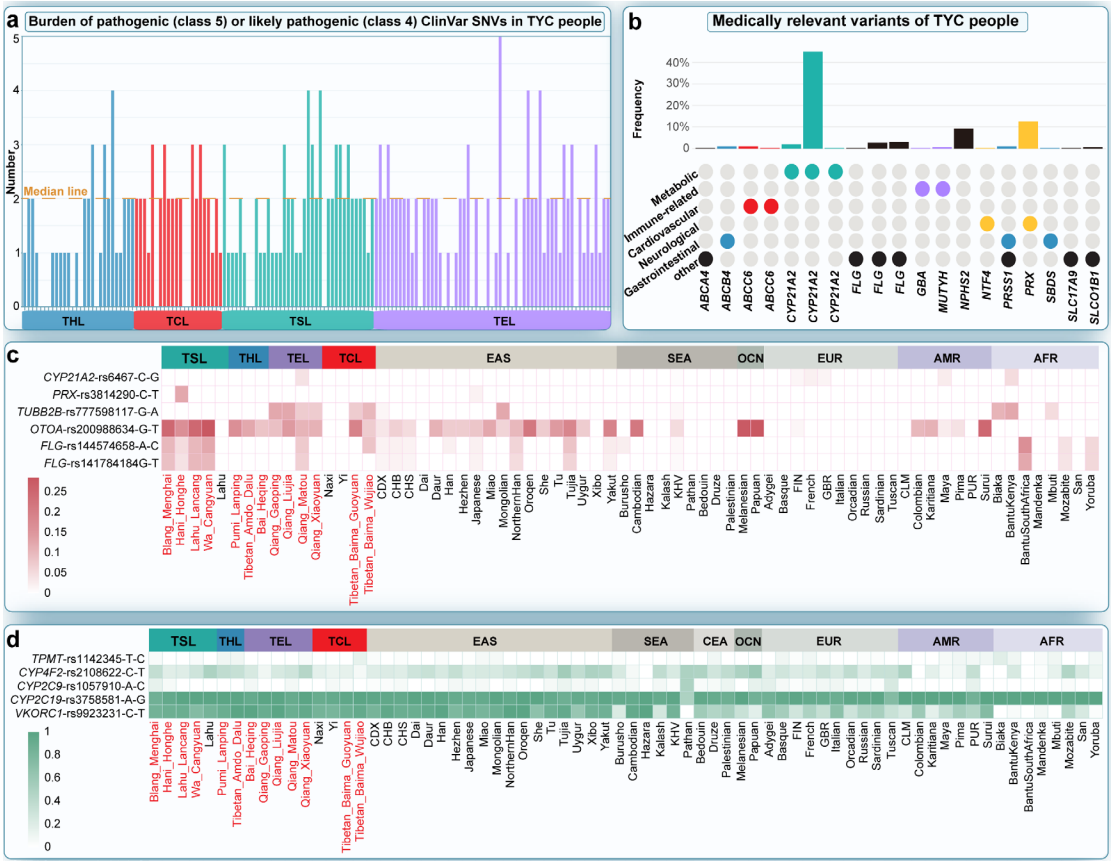

Fig. 4 | Carriers of pathogenic/likely pathogenic (P/LP) variants and pharmacogenetic variants in TYC patients.

(**a**) The burden of P (class 5) or LP (class 4) ClinVar SNVs per individual in the TYC population. (**b**) Juxtaposition of P/LP variants and the frequency of potential P/LP variants reported in genes from the American College of Medical Genetics and Genomics (ACMG) SF v3.2 list in TYC populations. (**c**) Frequency distribution of variants in genetically distinct TYC populations and worldwide populations from the HGDP and 1KGP resources, which are predicted to be P/LP according to ClinVar. For each population, the cell graphs depict the frequency of disease-associated alleles. (**d**) Population-specific frequencies of adverse drug reactions are predicted from the allele frequency distribution of known variants associated with drug response. EAS, East Asian; SEA, Southeast Asian; CEA, Central Asian; OCN, Oceanian; EUR, European; AMR, American; AFR, African.

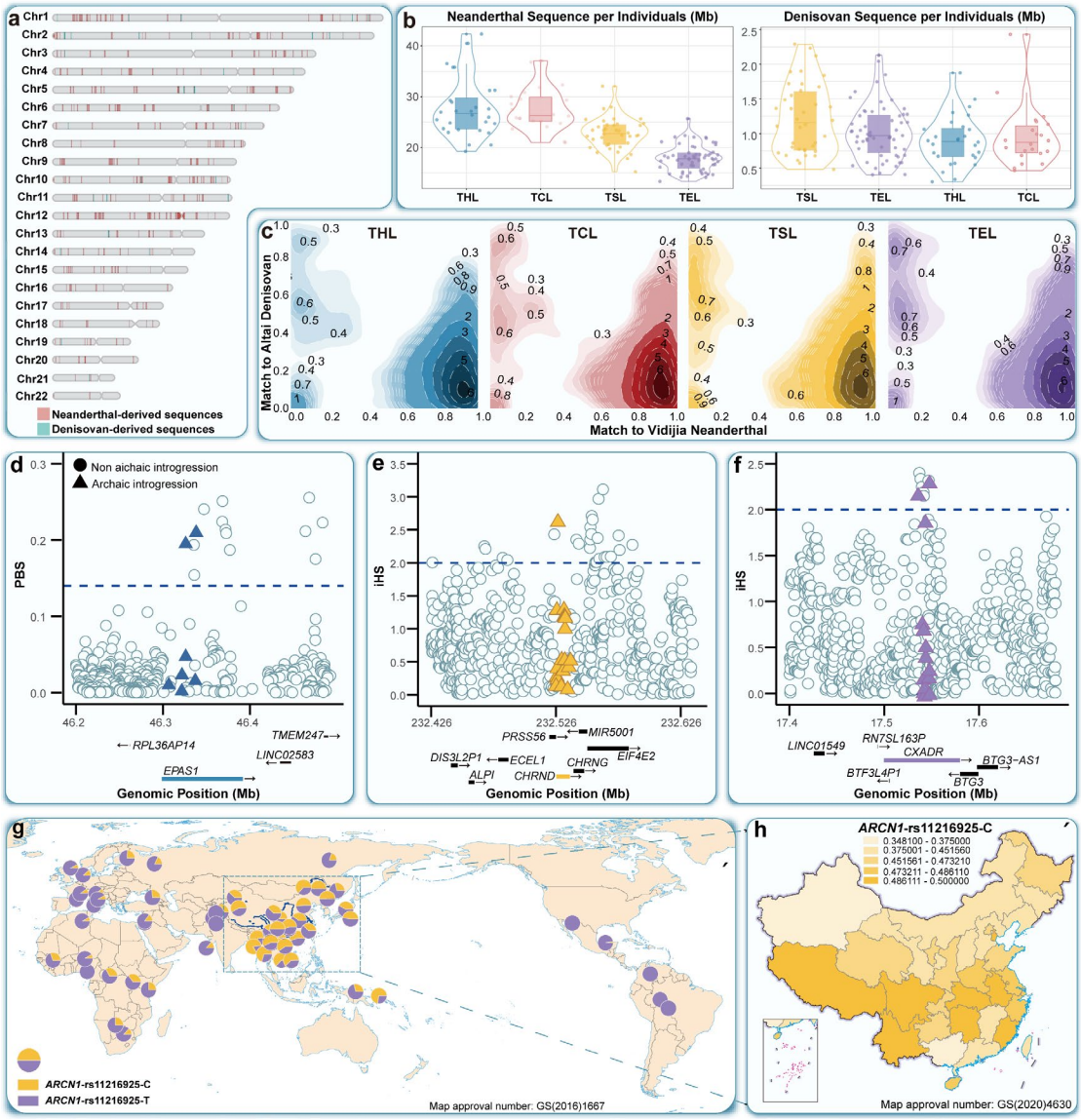

Fig. 5 | Genetic profiles and phenotypic effects of introgressed sequences from the Neanderthal and Denisovan populations.

(**a**) Density plot showing the distribution of introgressed sequences across each chromosome. Neanderthal-derived sequences are represented in red, whereas Denisovan-derived sequences are shown in blue. (**b**) Distribution of archaic-introgressed segment lengths on an individual basis. (**c**) Contour plot depicting the genome-wide match rate with archaic individuals in putatively introgressed segments. The x-axis represents the match rate to Neanderthals, and the y-axis represents the match rate to Denisovans. THL, TCL, TSL, and TEL are represented in blue, red, yellow, and purple, respectively. (**d-f**) Variants likely introgressed from Denisovans in the *EPAS1*, *CHRND*, and *CXADR* loci are associated with regulated stress erythropoiesis during hypoxia, the acetylcholine receptor signaling pathway, and cardiac muscle cell adhesion, respectively. The gray circles represent SNPs located within 100 kb upstream and downstream of the target gene, the colored triangles denote Denisovan-derived segments within the target gene, and the blue dashed line highlights regions that underwent positive selection. (**g**) Allele frequency of rs11216925, derived from Neanderthals in *ARCN1*, across global populations from the 10K_CPGDP, HGDP, and 1KGP genomic sources. (**h**) Frequency of the ancestral allele C, derived from Neanderthals, among Chinese individuals in the Huaxi Biobank. The base map was officially approved with the numbers GS (2016)1667 and GS(2020)4630 from the website http://bzdt.ch.mnr.gov.cn/.