Title: Whole-genome sequencing pilot of the Central Asian Genomic Diversity Project reveals distinct histories, adaptation, and introgression
Short title: Pilot whole-genome sequencing in Central Asia

Authors: Guanglin He[1,2,3]†, Haoran Su[1,3], Qiuxia Sun[1,2,3,10], Renkuan Tang[2,10], Qingxin Yang[1,3], Lintao Luo[1,2,3,10], Jie Zhong[1,3], Zhaxylyk Sabitov[5,6], Jing Cheng[1,3], Fengxiao Bu[1,3], Yu Lu[1,3], Chao Liu[7], Huijun Yuan[1,3], Lan-Hai Wei[4]†, Maxat Zhabagin[8,9]†, Mengge Wang[1,2,3,10]†

Affiliations
[1]Institute of Rare Diseases, West China Hospital of Sichuan University, Sichuan University, Chengdu, 610000, China
[2]Department of Forensic Medicine, College of Basic Medicine, Chongqing Medical University, Chongqing, 400331, China
[3]Center for Archaeological Science, Sichuan University, Chengdu, 610000, China
[4]School of Ethnology and Anthropology, Inner Mongolia Normal University, Hohhot, 010022, China
[5]Research Institute for Jochi Ulus Studies, Astana, 010000, Kazakhstan
[6]Kazakh Historical Society, Astana, 010000, Kazakhstan
[7]Anti-Drug Technology Center of Guangdong Province, Guangdong, 510000, China
[8]National Center for Biotechnology, Astana, 010000, Republic of Kazakhstan
[9]Nazarbayev University, Astana, 010000, Republic of Kazakhstan
[10]Human Genetics and Forensic Genomics Research Institute, College of Basic Medicine, Chongqing Medical University, Chongqing, 400331, China
†Correspondence: guanglinhe@wchscu.edu.cn (G.H.); mzhabagin@gmail.com (M.Z.); ryan.lh.wei@foxmail.com (L.W.); menggewang2021@163.com (M.W.)

Abstract
Central Asians are underrepresented in genomic research, limiting insights into their history and medical risk. We established the Central Asian Genomic Diversity Project and sequenced whole genomes of 166 individuals from 20 Central Asian and Afghanistan Hazara groups. Analyses reveal marked differentiation driven by varying West/East Eurasian ancestry; Tajiks align with West Eurasians, Dungans with East Asians. We identify four geographically structured Turkic-related clusters, two Indo-European clines, and long-range migration events, including Siberian links in Hazaras and Sino-Tibetan ties in Dungans. Admixture dates cluster ~650–1,000 years ago, coinciding with the Song–Yuan era and Mongol expansion. Variant discovery reveals distinct distributions of medically relevant variants and group-specific selection signals in metabolic, immune, and neurological pathways. Subsistence practice shifts correlate with trait-associated variation. Neanderthal-like and Denisovan-like segments show group-specific associations with immunity, psychiatric risk, drug metabolism, and type 2 diabetes. These data clarify Central Asian evolutionary history and inform precision medicine.

Teaser
Whole-genome data from 20 Central Asian groups reveal recent east–west admixture, selection signals, and medical risk variants.

MAIN TEXT
Introduction
The Human Genome Project transformed human genomics, enabling discoveries in demographic processes, adaptive histories, archaic introgression patterns, and the genetic basis of traits and diseases (*1-5*). Since the release of the draft genome, advances in sequencing and computational

methods, along with telomere-to-telomere and pangenome references, have allowed comprehensive analyses of genomic variation across ethnolinguistically diverse populations (*6-9*). However, most large-scale genomic research, like the UK Biobank and TOPMed Program, still focuses on individuals of European descent, which limits representation, hampers mechanistic understanding, and exacerbates health disparities (*10-16*). In response, large-scale initiatives across-continent populations have prioritized non-European cohorts, including the H3Africa, Simons Genome Diversity Project, GenomeAsia 100K, SG10K, and Mexican Biobank (*17, 18*) (*3-5*), in addition to regional resources from Oceania, Pakistan, Japan, and China, such as BioBank Japan, the China Kadoorie Biobank, Westlake BioBank, NyuWa, ChinaMAP, and the 10K Chinese People Genomic Diversity Project (10K_CPGDP) (*7, 10, 19-23*). Collectively, these efforts expand the global catalog of human variation and refine models of demographic history and disease origins.

Central Asia, situated as the Eurasian heartland, spans from Russia in the north to Iran and Afghanistan in the south and from the Caspian Sea in the west to China and Mongolia in the east. This region has historically served as a crossroads for the movement of people, languages, goods, and ideas between East and West Eurasia (*24-26*). Despite their pivotal role in ancient trans-Eurasian cultural exchanges, Central Asian populations remain underrepresented in human genomic sequencing efforts and medical genome research (*11*). Various demographic events have shaped the ethnolinguistic diversity of Central Asia. Anatomically modern humans are believed to have first occupied the region approximately 50 to 40 thousand years ago (kya). During the Paleolithic era, Central Asia was inhabited by a distinctive population known as Ancient North Eurasian (ANE), which contributed to the gene pool of Eurasia and the Americas through multiple migration events since the Upper Paleolithic (*25, 27*). To investigate the relationships among cultural practices, language distributions, and population movements, Narasimhan *et al*. analyzed whole-genome ancient DNA data from 523 individuals, primarily from Central Asia and northernmost South Asia. Their findings revealed that Neolithic individuals from Turan (present-day Turkmenistan, Uzbekistan, Tajikistan, Afghanistan, and Kyrgyzstan) and the central and southern Steppe (present-day Kazakhstan) presented ancestry related to West Siberian hunter-gatherers (WSHG), which were derived primarily from ANE and Eastern European hunter-gatherers (EEHG) (*26*). This WSHG-related ancestry was prevalent in eastern Turan and Kazakhstan before the arrival of the Yamnaya steppe pastoralists.

From the Bronze Age onward, gene flow from Western Steppe herders (WSH) and Iranian/Anatolian farmers transformed Central Asian populations and was linked to the spread of Indo-European languages (*26-28*). Although horse domestication is key to the steppe hypothesis, genomic data show that Botai individuals in northern Kazakhstan gained most of their ancestry from an ANE-related hunter-gatherer population, which is different from western steppe pastoralists, with additional input from ancient East Asians (AEA; Baikal_EN) (*27, 28*). Formal models suggest that Bronze Age Turan groups had mainly Iranian farmer-related ancestry, along with contributions from Anatolian farmers and WSHGs, and a small Andamanese hunter-gatherer (AHG) component (*26*). Bactria Margiana Archaeological Complex (BMAC) individuals had approximately 60–65% Iranian and 20–25% Anatolian farmer ancestry, plus approximately 10% WSHG and 2–5% AHG; steppe-derived ancestry was not present before 2100 BCE and only appeared in outliers dated roughly between 2100 and 1700 BCE (*26*). By the Middle–Late Bronze Age, steppe-related ancestry increased in Bactria communities, reflecting the growing influence of pastoralists (*26*). Since the Iron Age, migrations from southern Siberia and East Asia, linked to proto-Mongolic and proto-Turkic expansions, further shaped Central Asian gene pools (*25, 29, 30*). Analyses of 137 ancient genomes revealed highly structured Iron Age Scythians who mixed with eastern steppe pastoralists to form the Xiongnu confederations, which expanded westward and eventually contributed to the Huns (*29*). Throughout these periods, steppe ancestry increased

without entirely replacing earlier Iranian and Anatolian farmer-related ancestors (*26, 30*), resulting in widespread admixture and substantial East Asian ancestry across the region.

Subsequent cultural and political formations intensified this complexity (*25, 27-29*). From the fifth century CE, the Turkic, Uyghur, Khitan, and Mongol empires succeeded the Xiongnu/Hunnic confederations and propelled large-scale mobility across Eurasia; genetic, archaeological, and historical evidence points to a profound impact of Mongol westward expansion on present-day Central Asians (*29, 31, 32*). Today, the region includes Turkic/Mongolic speakers (Kyrgyz, Kazakh, Uyghur, Uzbek, Tatar, Karakalpak, Turkmen, Mongolian, and Kalmyk) and Indo-European speakers (Tajik, Russian, Hazara, Armenian, and Ukrainian), alongside groups with distinct linguistic affiliations (e.g., Dungan, Korean, and Chechen). Early low-density marker studies suggested exceptionally high diversity in Central Asia (*33*). More recent genome-wide single-nucleotide polymorphism (SNP) analyses revealed recent admixture and extensive gene flow from European, West Asian, East Asian, and South Asian sources (*27, 32, 34, 35*). For example, Turkic speakers show variable Asian ancestry traced to Mongolia and southern Siberia (*32*); Indo-European-speaking groups such as Yaghnobis and Tajiks have retained substantial continuity since the Iron Age, with limited recent influx (*36*). Contemporary Tajiks derive primarily from BMAC and Andronovo-related groups, with highland Tajiks carrying additional ancestry related to the Tarim mummies; Kyrgyz people's West Eurasian ancestry traces mainly to historical Xinjiang individuals (*24*).

Despite these advances, most studies rely on partial genomic data and small sample sizes, and there is no publicly available high-quality whole-genome sequencing (WGS) data from a large-scale population-level cohort of Central Asian and Afghanistan Hazara (CAAH) populations. Furthermore, the extent to which complex ancient population movements and cultural exchanges influenced the genomic diversity of ethnolinguistically diverse Central Asians remains unknown. This gap hampers a comprehensive understanding of genomic diversity and the contributions of various Eurasian ancestries to Central Asian heritage and adaptation. To address this, we generated WGS data for 166 individuals from nine ethnic groups (Kyrgyz, Uzbek, Uyghur, Turkmen, Karakalpak, Karluk, Tajik, Hazara, and Dungan) across Central Asia and Afghanistan as part of the pilot phase of the Central Asian Genomic Diversity Project (CAGDP). This study aimed to identify gaps in the current genomic landscape. We extensively analyzed the demographic and adaptive history of ethnolinguistically diverse Central Asians by combining publicly available WGS and genome-wide SNP data from both ancient and modern Eurasians (Fig. 1, A and B, and table S1). Our results provide insights into the evolutionary history and local adaptation of present-day Central Asians and improve our understanding of the genetic interactions between West and East Eurasians.

Results

Variant discovery, genetic structure, and population history of Central Asian people

We sequenced the genomes of 166 individuals from 20 CAAH populations, with an average coverage of 13.66×. This analysis identified 22,994,949 SNPs and 2,952,662 small insertions or deletions (InDels) across 22 autosomes and the X chromosome. Of these, 2,300,814 SNPs (10.0%) and 834,296 InDels (28.3%) were newly identified and absent from dbSNP version 156. Most of these unreported variants were extremely rare: 89.4% were singletons, 9.1% had a minor allele frequency (MAF) less than 0.01, and 1.2% had an MAF between 0.01 and 0.05. Among the 22,538,640 biallelic SNPs observed, 40.2% were singletons, 15.1% had an MAF less than 0.01 (excluding singletons), and 14.7% had an MAF between 0.01 and 0.05. We calculated a transition versus transversion ratio (Ti/Tv) of 2.05 based on 21,608,632 biallelic autosomal SNPs.

We conducted a principal component analysis (PCA) to assess the genetic relationships between newly sequenced CAAH samples and Eurasian reference populations via the WGS dataset. These findings indicated that CAAH individuals formed a distinct genetic structure

between the West Eurasian clines and East Eurasian clines. Within this structure, Tajiks were more closely related to West Eurasians, whereas Dungans clustered with East Eurasians (fig. S1). To explore finer-scale population differences, additional ancient and modern Eurasian population data from the Human Origins (HO) dataset and published microarray-based genomic sources were integrated (*27, 32, 37*). Non-Dungan groups aligned along the West Eurasian-related cline, whereas Dungans were situated on the northern branch of the East Eurasian-related cline (fig. S2). Specifically, Tajiks and Karluks presented close genetic links to Indo-European and Turkic-speaking populations from the Caucasus. Uzbeks and Turkmens were spread among Turkic groups in southwest Central Asia, whereas the Kyrgyz, Uyghur, Karakalpak, and Hazara people overlapped with Turkic groups from northern and eastern Central Asia and China's Xinjiang. The Dungans displayed genetic connections with Mongolic, Tibeto-Burman, and Turkic-speaking populations from China. Further analysis of the genetic coordinates for CAAH, including linguistically similar groups, revealed a transparent east–west genetic gradient among Turkic-speaking groups across Eurasia. Tuvinians, Todzins, and Tofalars from southern Siberia, along with Dolgans and Yakuts from North Asia, were genetically close to Tungusic groups of the Russian Far East and Mongolic groups from Mongolia and Siberia. Conversely, Turkic groups from the Caucasus, such as Balkar, Karachai, Kumyk, and Azeri, presented closer genetic ties with Indo-European groups from West Eurasia (Fig. 1C). Indo-European speakers were distributed along two main genetic gradients: one included Armenian, Kurd, Tajik, and Yaghnobi individuals from southern Central Asia, the Caucasus, and West Asia; the other encompassed Russian, Ukrainian, and German individuals from Eastern Europe (Fig. 1C). With respect to ancient reference populations (fig. S2), non-Dungan groups overlapped with or were closely related to individuals from the Iron Age Xinjiang and historical Kazakhstan. The Dungan population had strong genetic links to ancient populations from the Yellow River basin (YRB) and the Qinghai–Xizang Plateau. Notably, Kyrgyz, Uyghur, Karakalpak, and Hazara individuals shifted toward ancient populations with less West Eurasian ancestry from the Mongolian Plateau, indicating increased shared genetic drift with East Eurasian-related ancestries. Conversely, Tajiks and some Uzbek/Turkic individuals were positioned closer to Iranian/Anatolian farmer-related populations, Russian steppe groups, ancient populations from Turan (excluding Kyrgyzstan), and Middle to Late Bronze Age populations from Kazakhstan, suggesting greater allele sharing with West Eurasian-related ancestries, especially Iranian farmers, Anatolian farmers, and WSH-related ancestries. These individuals also demonstrated close genetic ties to ancient populations from Pakistan. Additionally, Tajik, Karluk, Uzbek, and Turkmen individuals presented genetic affinities with ancient populations from Kyrgyzstan.

Patterns of genetic differentiation among CAAH groups were confirmed through model-based ADMIXTURE analysis (figs. S3 and S4). PCA and ADMIXTURE findings revealed that geographically separated, ethnolinguistically uniform groups such as Tajik, Turkmen, Hazara, and Kyrgyz shared similar ancestry profiles (fig. S4). These groups were labeled accordingly. At K = 13 (where cross-validation plateaued, fig. S3), Kyrgyz and Karakalpak derived most of their ancestry (> 10% each) from ancient Northeast Asian (ANA; Boisman_MN), EEHG (Yaroslavl_VolosovoLyalovo_N), and Iranian farmer-related sources (TepeAbdulHosein_N), as well as additional contributions (2–10% each) from ancient northern East Asian (ANEA; Lhasa1k), Neo-Siberian (Nganasan), Japanese, southern East Asian (Igorot), Anatolian farmer (Anatolia_N), and ANE-related (Tarim_EMBA1) ancestries (Fig. 1, D to F). Uyghur, Hazara, Uzbek, and Turkmen presented these nine components, with Uyghur and Hazara being richer in Iranian farmer, ANA, EEHG, and ANEA ancestries, whereas Uzbek and Turkmen presented higher levels of Iranian farmer, EEHG, ANA, and Anatolian farmer ancestries. Karluk's primary ancestry consisted of the Iranian farmer and EEHG, with minor contributions from ANA, Anatolian farmer, ANEA, ANE, Japanese, Neo-Siberian, SEA, and ancient Xinjiang populations (Wutulan). Tajiks' main ancestry was from Iranian farmers, EEHG, and Anatolian farmers, with

smaller proportions from ANE, ANA, ANEA, Wutulan, and Japanese. Dungan had East Asian (ANEA, SEA, Japanese) and ANA ancestries, along with Iranian farmer and EEHG inputs. Outgroup f_3-statistics indicated that the Kyrgyz, Uyghur, Karakalpak, and Hazara populations shared more alleles with the Altaic groups in northern East Asia and Siberia, as did ancient populations from the Amur River basin (ARB), West Liao River (WLR), and YRB (figs. S5 and S6). Uzbek and Turkmen shared greater alleles with the Tungusic, Turkic, and Uralic groups from Siberia; Karluk shared more alleles with the Siberian and European Uralic groups, Russians from Eastern Europe, and several Turkic and isolated Siberian groups. Tajiks shared more alleles with European Indo-European and Uralic groups, as well as with Caucasian-speaking populations from the Caucasus. These groups also shared excess alleles with early and middle Bronze Age Tarim and steppe-related populations from Kazakhstan and southern Siberia. Dungan presented more allele sharing with East Asian (notably Sinitic) and ancient YRB groups than others did, indicating different patterns of admixture. TreeMix analysis supported these allele-sharing patterns (fig. S7).

Geography-related genetic differentiation among Turkic groups across Eurasia

Turkic groups are widely distributed across Central Asia and surrounding regions, yet their population structures, admixture profiles, and genetic origins are not fully understood. Large-scale genomic data from 54 geographically distinct Turkic-speaking populations across Eurasia, combined with advanced analytical methods, provide valuable insights into their deep history. PCA and ADMIXTURE analyses revealed clear genetic differences, identifying four major clusters aligned with geography (Fig. 1, C and F, and figs. S2 and S4). The East Asian cluster included mainly Salar and Yugur populations from Northwest China (excluding Xinjiang). The Siberian cluster consisted of, but was not limited to, the Altaian, Khakass, Shor, Tubalar, Tofalar, Tuvinian, and Yakut populations from southern Siberia and the Russian Far East. The Central Asian cluster included Kazakh, Kyrgyz, Karakalpak, and Uyghur populations from northern and eastern Central Asia, as well as Xinjiang, China. The West Eurasian cluster included Turkmen, Uzbek, Nogai, Balkar, Bashkir, Kumyk, and Tatar populations from the Caucasus, Eastern Europe, West Asia, western Siberia, and southwestern Central Asia (Fig. 1C). Pairwise Welch's t-tests of the first two PCs confirmed the presence of geography-related substructures within the Turkic groups (Fig. 2A). The TreeMix-derived topology supported distinct population structures and revealed complex gene flow among groups (fig. S8). To explore the factors influencing the observed genetic differences, correlations between geographic factors (latitude and longitude) and genetic features (PC coefficients and ancestral component proportions) were examined (Fig. 2B). PC1 was strongly negatively correlated with longitude ($r = -0.89$, $p < 0.001$), whereas PC2 was strongly positively correlated with longitude ($r = 0.77$, $p < 0.001$) and moderately positively correlated with latitude ($r = 0.59$, $p < 0.001$). These patterns suggest pronounced genetic differences between eastern and western, as well as northern and southern, Eurasian Turkic groups. The percentage of Neo-Siberian-related ancestry was strongly positively correlated with latitude ($r = 0.70$, $p < 0.001$), whereas ANA- and ANE-related ancestries showed weak positive correlations ($r < 0.30$, $p < 0.05$). Conversely, moderate negative correlations with latitude were observed for ANEA ($r = -0.45$, $p < 0.001$), SEA ($r = -0.49$, $p < 0.001$), and Iranian farmer-related ancestry ($r = -0.52$, $p < 0.001$), with a weaker negative correlation for Japanese ancestry ($r = -0.31$, $p < 0.05$). With respect to longitude, ANA-related ancestry had a strong positive correlation ($r = 0.72$, $p < 0.001$). Moderate positive correlations were also found for East Asian and Neo-Siberian-related ancestries ($0.40 < r < 0.60$, $0.001 < p < 0.01$). In contrast, the Anatolian farmer- and EEHG-related components showed extremely strong negative correlations with longitude ($r \leq -0.80$, $p < 0.001$), and Iranian farmer-related ancestry also displayed a strong negative correlation ($r = -0.67$, $p < 0.001$). Interestingly, the relationships between geographic distance and genetic differences among the Altaic-speaking groups (including the Turkic, Mongolic, Tungusic,

Koreanic, and Japonic groups) differed slightly from those within the Turkic groups (fig. S9). Among Altaic speakers, ANE-related ancestry showed a weak negative correlation with longitude ($r = -0.37$, $p < 0.001$), whereas no substantial correlation with latitude was observed within Turkic groups for this component. We finally confirmed the strong correlation between genetic and geographic variation among Turkic groups through Procrustes analyses (Fig. 2C).

We performed admixture modeling on all Turkic groups to reconstruct their population structure via four ancestral sources. These sources were prominent in Gagauz from Eastern Europe, Tofalar in Siberia, Tubar in southwestern Siberia, and Xunhua Salar in East Asia (Fig. 2D). To explore the fine-scale genetic structure, the genomes from 54 populations were phased and analyzed via the fineSTRUCTURE and IBD methods (Fig. 2, E to H, and figs. S10 to S12). The haplotype dendrogram showed that five major clusters aligned with geographic regions (Fig. 2E), mostly matching the regional PCA results (Fig. 1C). The Siberian cluster is split into southwestern Siberian (Tubalar and Shor) and southern Siberian–Russian Far Eastern (Altaian, Yakut, Tuvinian, Tofalar, Khakass, Dolgan) subgroups. IBD sharing revealed extensive gene flow within the past 1,500 years among the Turkic and Mongolic/Tungusic groups in Siberia (figs. S10 and S11A), alongside ancient links across Eurasian Turkic groups (fig. S11B). Longer IBD segments indicated gene flow between the easternmost Turkic groups (Dolgan and Yakut) and neighboring Turkic, Tungusic, and Mongolic groups, mostly from ancient events following a distance decay pattern (Fig. 2F, and fig. S12, A to D). We detected excess IBD sharing between West Eurasian Turkic groups and geographically proximate Indo-European groups, extending to Indo-European groups in Central Asia and to Turkic and Tungusic groups in Siberia, a pattern primarily driven by deeper ancestral links (fig. S12, E to H).

Long-distance migration of Dungans and Hazaras shaped their distinct genetic makeup

The genetic relationships among geographically separated populations mostly follow an isolation-by-distance model. However, strong genetic links among distant Eurasian groups provide direct evidence of long-range migrations. Clustering patterns from PCA, ADMIXTURE, and outgroup f_3-statistics revealed that Dungan and Hazara did not have close genetic ties to nearby Central Asian groups. Instead, both had closer genetic connections to East Eurasian populations (Fig. 1, C and F, and figs. S1, S2, and S4 to S7), a pattern further supported by pairwise F_{ST} genetic distances (fig. S13). Specifically, Dungan shared more alleles with East Asians, especially Sino-Tibetan speakers, than with neighboring Central Asians. The Hazara population displayed admixture profiles like those of the Uyghurs, with additional affinity to Altaic groups from northern East Asia and Siberia. In Dungan-related Sino-Tibetan groups, the proportions of inferred ancestries, excluding ANEAs, SEAs, and Neo-Siberians, increased with latitude ($0.44 \leq r \leq 0.85$, $p < 0.01$), whereas SEA-related ancestries showed a weak negative correlation ($r = -0.31$, $p < 0.05$) (fig. S14A). These ancestries, excluding ANAs, SEAs, and Japanese individuals, were negatively correlated with longitude ($0.42 \leq |r| \leq 0.63$, $p < 0.01$). Conversely, the SEA ($r = 0.52$, $p < 0.001$) and Japanese-related ($r = 0.54$, $p < 0.001$) components were positively correlated with longitude. In Hazara-related Indo-European groups, ancestries linked to East Asians, ANAs, ANEs, and Iranian farmers were negatively correlated with latitude ($0.49 \leq |r| \leq 0.86$, $p < 0.05$), whereas these components, except for ANAs, were positively associated with longitude ($0.53 \leq r \leq 0.87$, $p < 0.05$) (fig. S14B). To explore haplotype-level relationships, IBD sharing between Dungan/Hazara populations and various Eurasian populations was estimated (Fig. 2, G and H, and fig. S14, C to F). Extensive ancient connections were found between the Dungan and East Asian populations, particularly the Sino-Tibetan groups (Fig. 2G, and fig. S14, C and D). Hazara showed excess IBD sharing with populations that are geographically and linguistically different, mainly Mongolic and Turkic groups in northern East Asia, followed by Turkic groups in Central Asia and Altaic groups in Siberia, reflecting deep genetic ties (Fig. 2H, and fig. S14, E and F). Overall, the patterns of allele and haplotype sharing of Dungan with Sino-Tibetan groups in East

Asia and Hazara with Altaic groups in northern East Asia and Siberia, along with the prominent East Asian ancestry in Dungan and substantial East Eurasian ancestry in Hazara, suggest that long-distance migrations of ancestral groups related to these populations have shaped their genetic makeup.

Admixture modeling and admixture time estimation of CAAH

We initially applied admixture f_3-statistics to identify potential source populations for CAAH groups. For all CAAH populations except Karluk and Dungan, the most negative f_3-values appeared with reference pairs combining one eastern and one western Eurasian source. This pattern strongly suggests that these CAAH groups arose from admixture between deeply diverged eastern and western Eurasian lineages. In contrast, the distinct admixture signals in Karluk and Dungan imply a more complex genetic history (tables S2 and S3). Our analysis revealed that reference pairs involving Nanai and Ulchi produced more markedly negative values in Kyrgyz than those involving other East Eurasian sources. For Uyghur, Uzbek, and Hazara, reference pairs with Han Chinese, Japanese, Nanai, and Ulchi produced more negative f_3-values. Additionally, Uyghur and Uzbek populations presented more negative f_3-values when ANA-related ancestors served as East Eurasian sources. Turkmen and Karakalpak presented more negative f_3-scores, with Nanai, Ulchi, Nivkh, and ANA-related ancestries as references (table S2). Notably, pairs involving Anatolian and Iranian farmer-related ancestries generated more negative f_3-scores in Tajiks than those with other West Eurasian sources (table S3). To explore genetic diversity among geographically separated CAAH groups, we performed f_4-statistics of the form f_4(CAAH1, CAAH2; Reference, Mbuti). The population stratification patterns from this symmetric analysis matched the descriptive results, showing that genetically distinct CAAH populations share alleles differently from the West and East Eurasian reference groups (fig. S15). Compared with other CAAH groups, Dungan displayed greater genetic drift with AEA- and ANA-related ancestries, as well as with Scythian-related ancient Central Asians (ACAs) characterized by substantial East Asian ancestry (*29*). This was followed by Kyrgyz, Karakalpak, Hazara, and Uyghur, which also presented increased allele sharing with AEAs, ANAs, and East Asian-affinity ACAs. Compared with Tajiks and Turkmen, Uzbeks presented stronger genetic links with AEAs and East Asian-affinity ACAs, along with more genetic drift with ANAs and ANEAs, than Karluk did. Moreover, Turkmen and Karluk shared more alleles with AEAs, ANAs, and East Asian-affinity ACAs than Tajik did. Conversely, the allele-sharing patterns between CAAH groups and Anatolian farmers, Iranian farmers, and steppe pastoralist ancestors differed markedly from those involving AEAs, ANAs, and East Asian-affinity ACAs (fig. S15).

To investigate the ancestry profiles of CAAH groups, qpAdm was used to determine whether the target population could be modeled as a mixture of specific source populations and to estimate their respective ancestry proportions. The best-fitting model was a two-way admixture involving Tajikistan_Ksirov_Kushan (30%–86%), a historic nomadic group descending from Hajji_Firuz_C, BMAC, and Steppe_LBA-related sources, and Xianbei_IA (14%–70%). Most CAAH populations fit this model well, except for Dungan and Karakalpak (Fig. 3A). Dungan was best modeled, with only 8.6%–9.3% Tajikistan_Ksirov_Kushan ancestry, with the remainder (90.7%–91.4%) from ANEAs. Karakalpak can be simplified into a two-way admixture from West and East Eurasian steppe pastoralists, approximately 42% Alan, which likely descended from Sarmatians and modern Huns, and 58% Xianbei_IA. The haplotype-based SOURCEFIND and fastGLOBETROTTER methods supported these results, revealing a detailed ancestry structure within CAAH groups. SOURCEFIND estimates revealed variable contributions of haplotypes from the French, Kalash, Bedouin, Yakut, and northern Han populations (fig. S16). FastGLOBETROTTER analyses suggested a single admixture event for most CAAH groups, except Tajik, which experienced multiple events (Fig. 3B and fig. S17). West Eurasian ancestry, linked to French ancestry, made up 32%–45% of the Kyrgyz, Karakalpak, Hazara, and Uyghur

genomes and over 50% of the Turkmen, Uzbek, and Karluk genomes (57%–62%). The remaining genetic component was derived mainly from Yakut, indicating East Eurasian influence. Dungan was modeled as approximately 87% Northern Han and 13% French. Tajik ancestry appeared to involve two distinct admixture events, with varying contributions from French, Bedouin, Kalash, Northern Han, and Yakut populations (fig. S17B).

Admixture dating was conducted via MALDER and fastGLOBETROTTER to analyze east–west gene flow. MALDER-based estimates revealed a single admixture event involving a West Eurasian source (French) and an East Eurasian source (Yakut or Northern Han), which was estimated to have occurred approximately 23–34 generations ago (Fig. 3B). The FastGLOBETROTTER results for Turkmen, Karluk, Uyghur, and Karakalpak were consistent, dating admixture to approximately 22–25 generations ago. More recent admixture events were suggested for Uzbek, Hazara, Kyrgyz, and Dungan, which were estimated to have occurred approximately 22–26 generations ago (Fig. 3B). In Tajik, two distinct admixture pulses were identified, with the most recent pulse occurring approximately 9 generations ago and an older event occurring approximately 50 generations ago. SMC++ was used to estimate the effective population size (Ne), which revealed that Karluk had a smaller Ne than the other CAAH groups did. While most CAAH populations experienced marked growth over the past 10,000 years, peaks in Ne were observed in Karluk, Uyghur, and Hazara approximately 2–3 kya, followed by declines (Fig. 3C). The divergence times were approximately 20.8–15.3 kya between East Eurasians (Northern Han) and CAAH, approximately 2.6–7.2 kya between West Eurasians (French) and CAAH, and approximately 1.9–3.3 kya among CAAH groups (Fig. 3D and fig. S18A). Notably, the proportion of West Eurasian ancestry in CAAH populations was positively correlated with the divergence time from East Eurasians and negatively correlated with the divergence time from West Eurasians (fig. S18, B and C). We reconstructed deep demographic models of four representative Central Asian lineages based on coalescence simulations. We compared these models with their East Asian origin and other candidate models of western Eurasian origin, considering both symmetric and asymmetric gene flow events (Fig. 3E). The best-fit model, which yielded the highest likelihood, indicated that East and West Eurasian lineages diverged ~37.4 kya [95% confidence interval (CI), 37.2–40.2 kya]. The Turkic-speaking Karakalpak and Kyrgyz shared a common ancestor with Dungan and Han Chinese, with divergence occurring around 7.6 kya (95% CI: 7.4–8.8 kya). The Karakalpak–Kyrgyz lineage experienced significant gene flow from the Tajik-related lineage. Divergence between the French and Tajik occurred ~2.0 kya (95% CI: 1.2–7.4 kya), between Dungan and Han Chinese ~1.8 kya (95% CI: 0.7–4.6 kya), and between Karakalpak and Kyrgyz ~0.6 kya (95% CI: 0.4–3.4 kya).

Sex-biased admixture landscape

To investigate sex-biased admixture in CAAH groups, we estimated and compared the proportions of source populations via autosomal, X-chromosomal, Y-chromosomal, and mitochondrial data. Differences between autosome and X-chromosome-based inferred ancestry proportions suggest sex-biased admixture, since males inherit only their mother's X chromosome. The western Eurasian genetic contribution ranged from 7.8% to 86.4% for autosomes and from 8.4% to 85.8% for X chromosomes (Fig. 3F). A substantial female bias toward West Eurasian ancestry was detected in Hazara ($Z = -0.43$; Student's t-test, $p < 0.05$), which is consistent with an ~71.4% enrichment of West Eurasian-related maternal lineages, such as haplogroups H, J, K, N1, T, U, and W (table S4). No notable sex bias was found in the other CAAH populations ($-0.20 < Z < 0.13$; Student's t-test, $p > 0.05$). However, several groups presented differences in West Eurasian paternal or maternal lineages compared with autosomal and X-chromosomal ancestry estimates (Fig. 3F and table S4). All Karluk individuals carried the Y-chromosome haplogroup L1a2a, which is common in South Asian populations. West Eurasian-related paternal lineages appeared in 65.0% of Kyrgyz, 35.3% of Karakalpak, and 30.0% of Dungan, whereas West

Eurasian maternal lineages were present in 88.9% of Karluk, 70.0% of Tajik, 55.0% of Uzbek, and 31.6% of Uyghur (table S4). Overall, no evidence suggests a simultaneous excess of West Eurasian paternal and East Eurasian maternal lineages or vice versa among Central Asians.

Genomic signatures of positive selection

Local biological adaptation can cause changes in allele frequencies (AFs) in admixed populations that differ from predictions based on admixture proportions (*38, 39*). To detect post-admixture adaptive signals in CAAH populations and evaluate their biological significance, deviations (AFd_e) between the observed (AF_{obs}) and expected allele frequencies (AF_{exp}) were calculated (table S5). Owing to the large-scale geographic sampling and the genetic differentiation within CAAH groups revealed by fineSTRUCTURE, analyses were performed separately for four subgroups: CAAH-West, CAAH-Mixed, CAAH-Kyrgyz, and CAAH-Dungan (fig. S19). Genes with substantial AFd_e variants were involved mainly in pathways related to metabolism, development, the immune response, and neural functions (fig. S20A). Additionally, approximately 5.00–6.86% of the variants with strong AFd_e across these groups presented evidence of hard selective sweeps, as indicated by integrated haplotype scores (|iHS| > 2; table S6). The most prominent signals of metabolic adaptation were in *EBF2*, which regulates adipogenesis, particularly in CAAH-West and CAAH-Mixed (Fig. 4, A and B). Several variants in the *EBF2* 3′UTR (e.g., rs13257257-G, rs17054430-C) had higher AF_{obs} (0.34–0.42) than expected post-admixture (0.16–0.18) and have been linked to altered brown adipogenesis, increased white fat storage, and a predisposition to inguinal hernia (*40*). In CAAH-West, variants in *FNDC1*, which is involved in cellular responses to hypoxia and positively regulates cardiomyocyte apoptosis, also showed remarkable frequency deviations, including rs294883-T, which was previously associated with coronary artery disease (*41*) (Fig. 4A). In CAAH-Mixed, two variants in *CARS2*, which influence the synthesis of enzymes involved in serine synthesis (rs7324648-A and rs373373-C), were notably enriched (*42, 43*) (Fig. 4B). Variants in *MTAP*, a key gene in polyamine metabolism, also presented strong selection signals (Fig. 4B and table S5). Notably, rs871024, which is linked to melanoma, was more common in Hazara and South Asian populations, with a marked increase in derived allele frequency (DAF) during the transition from hunter-gatherer to farming lifestyles, after which it stabilized during the nomadic period (fig. S20, B to D). In CAAH-Kyrgyz, variants in *CLDN10* (rs77008184, rs2095774, and rs17189719), which are essential for renal function and intestinal barrier integrity, show differentiation and are associated with serum uric acid and cystatin C levels (*42, 43*) (Fig. 4C). The DAFs of rs77008184 and rs17189719 decreased during the shift from hunting-gathering to farming and declined further with the expansion of pastoralism, whereas rs2095774 showed the opposite trend (fig. S21, A to C). In CAAH-Dungan, clusters of variants in *LINGO2*, previously linked to body mass index and fat distribution, displayed extended haplotype homozygosity (EHH) patterns indicative of local adaptation (Fig. 4, D and E, and figs. S21D and S22). These findings suggest that various CAAH groups have experienced unique metabolic adaptations driven by their distinct genetic backgrounds.

Post-admixture adaptation signatures related to immune pathways were concentrated in *TRAF1* and *PHF19*, which were identified in CAAH-West (Fig. 4F, fig. S23A, and table S5). These loci included alleles linked to allergic diseases (*PHF19*-rs10760123/rs6478484/rs10818483 and *TRAF1*-rs1930781), rheumatoid arthritis (RA; *TRAF1*-rs10435844/rs3761847, *PHF19*-rs1953126/rs881375/rs2072438), white blood cell count (*TRAF1*-rs2416804), and lymphocyte count (*PHF19*-rs7858209) (*44-47*). DAFs of *TRAF1*-rs2416804/rs10435844/rs3761847 and *PHF19*-rs2072438/rs7858209 declined before the agricultural era and again after the rise of nomadic pastoralism, whereas other loci showed opposite trends (fig. S23, B to L). EHH decay for these variants indicated strong positive selection on ancestral alleles, resulting in substantial frequency differences among CAAH groups and other continental populations (Fig. 4, G to N).

We further reconstructed the DAF evolutionary trajectory among spatiotemporally ancient Eurasian genomes and confirmed that the AF shifted during transitions to agriculture and nomadic pastoralism (*48*) (fig. S23M). The selection signals in CAAH-Dungan included variants in *LPXN*, *PPP4R1L*, and *GIMAP6*: *LPXN*-rs10896794, linked to inflammatory bowel disease; *PPP4R1L*-rs480143, associated with white blood cell count; and *GIMAP6*-rs62491814, related to C-reactive protein levels (*44, 49, 50*) (Fig. 4D and table S5).

Selection signals were also enriched at loci affecting neural traits. Notably, the variant rs4708181 in *FILIP1*, identified in CAAH-Mixed, CAAH-Kyrgyz, and CAAH-Dungan (Fig. 4, B to D), has been linked to cortical surface area (*51*). In CAAH-Mixed, strong selection signals were observed in *CHRNA3*, with rs112878080 linked to carbon monoxide diffusing capacity and nine additional variants associated with pulmonary traits, including the FEV1/FVC ratio and post-bronchodilator FEV1 (*52, 53*) (Fig. 4B and table S5); notably, the DAFs of rs112878080, rs138544659, and rs8040868 declined ~5,500 years ago (fig. S24, A to C). The signals were also concentrated in *HDAC9*, a gene implicated in brain morphology and neurological conditions (fig. S24D). Variants associated with cortical surface area, cortical thickness, and whole-brain restricted diffusion, as well as male-pattern baldness (*54-57*), excluding rs13242758 and rs2073963, showed dynamic DAF trajectories from ~9–6 kya, stabilizing following the regional transition to nomadic pastoralism (fig. S24, E to O). Another candidate, *SPON1*, encoding an axon-guidance protein, harbors the variant rs12575169, which is associated with adolescent idiopathic scoliosis (*58*), and rs2618516, which is linked to brain connectivity (*59*), both of which are under selection in CAAH-Kyrgyz. Furthermore, *TCF4*, a key regulator of neural development, exhibited strong selection signals in CAAH-Dungan (Fig. 4D and table S5). Variants associated with neuroticism, risk-taking, and mood-related phenotypes (e.g., rs2919451 and rs2924328) presented a marked increase in DAFs between ~9 and 7.5 kya, followed by another increase from ~5 to 3 kya and a pronounced decrease of approximately 3 kya (*60, 61*) (fig. S24, P and Q).

To further refine adaptive candidates, variants exhibiting high differentiation between East and West Eurasians were intersected with AFd_e signals. Variants with an $F_{ST} > 0.2$ were defined as highly differentiated variants (HDVs), yielding 369,407 candidates. Among these, those with $AFd_e > 0.1$ in specific CAAH subgroups were considered potentially adaptive (table S7). A strong selection signal was observed at *THOP1* in CAAH-West (Fig. 4A), a locus associated with testosterone and low-density lipoprotein cholesterol levels. EHH curves supported local adaptation at these variants (fig. S25, A to E). Another notable signal involved *VPS41* in CAAH-Mixed and CAAH-Kyrgyz, where rs6943746 and rs10274968 are linked to neuroticism and depression (table S7). Although most adaptive signals displayed a west Eurasian frequency bias in CAAH-Mixed and CAAH-Kyrgyz, HDVs with elevated AFd_e in CAAH-Kyrgyz exhibited an east Eurasian shift (Fig. 5, A and B). The DAF of *VPS41*-rs10274968 increased markedly between 9–7 kya, followed by a slight increase before stabilizing at ~4 kya (Fig. 5C). Additional adaptive signals involved genes associated with nervous system function, including *CTNNA2* (chronic pain) in CAAH-Mixed; *PLCL1* and *DPP10* (psychiatric disorders) in CAAH-Kyrgyz; and *PLCH1*, *PTPRD*, and *FGF14* (insomnia) in CAAH-Dungan (*62, 63*) (Fig. 4, B to D, and table S7). Immune-related loci are also strongly differentiated. Examples included *CSMD1*-rs1529316 (multiple sclerosis) in CAAH-Mixed and CAAH-Dungan, *SGCZ*-rs9886428 (IgG glycosylation) and *IL12B*-rs3213094 (psoriasis) in CAAH-Mixed, *TLR1*-rs5743614/rs6531663 (atopic dermatitis/allergic disease) in CAAH-Kyrgyz, and *SLC9A4*-rs1468788 (celiac disease) in CAAH-Dungan. *IL12B*-rs3213094 presented increased DAFs with the transition to agriculture but declined prior to the advent of nomadic pastoralism (fig. S25F). The DAF of *TLR1*-rs5743614 has decreased since ~6 kya, whereas that of rs6531663 has steadily increased over the same period (fig. S25, G and H). Notably, the DAF of *SLC9A4*-rs1468788 increased continuously over the past 10 kya (Fig. 5D). Furthermore, these immune-related variants presented obvious geographic variation in frequency (fig. S26). A diet-related signal associated with meat-based dietary

adaptation was observed at *CBX5*-rs4759074 in CAAH-Dungan, with a frequency peak ~6 kya preceding the emergence of pastoralism (Fig. 5E). Population-specific enrichment was also found for *CBX5*-rs17110109 (CAAH-Dungan) and *ITPR1*-rs13313995 (CAAH-Kyrgyz), both of which are linked to alopecia (*57, 64, 65*) (table S7). These *CBX5* variants presented marked AF differences between CAAH-Dungan and other CAAH groups, as well as across continental populations (Fig. 5, F and G).

Neanderthal and Denisovan introgression and the biological functions of archaic heritage

Ethnolinguistically diverse Eurasian and Oceanian populations have heterogeneous interactions with archaic groups, resulting in distinct introgression patterns and a persistent genetic legacy (*66, 67*). The extent of Neanderthal and Denisovan introgression in Central Asian populations remains only partly understood. We used IBDmix to identify archaic introgression segments (AISs) from the Altai Neanderthal genome. The distribution of Neanderthal-derived sequences in CAAH-West closely resembled that of European populations from the 1000 Genomes Project (1KGP), whereas patterns in CAAH-Mixed mirrored South Asians (Fig. 6A). In contrast, CAAH-Kyrgyz and CAAH-Dungan had the shortest Neanderthal AISs per individual, likely due to sample size effects (*68*). On average, each carried ~50.01 Mb of Neanderthal sequences, ranging from 52.95 Mb in CAAH-West and CAAH-Mixed to 39.23 Mb in CAAH-Kyrgyz. We identified 104 Neanderthal-introgressed regions absent from the 1KGP (fig. S27A). We identified numerous archaic introgressed genomic regions and genes unique to Central Asians, such as *GUSBP1*, *ASTN2*, *ARHGAP15*, *CYP39A1*, *MAN1A2*, and *BCL11A*, which directly affect their genetic determinants of diseases and physical traits. For example, a *BCL11A* segment is associated with hematopoietic cell development and hematological disorders. Additionally, fifty-four segments were specific to CAAH subgroups: 27 in CAAH-West, 18 in CAAH-Mixed, 4 in CAAH-Kyrgyz, and 5 in CAAH-Dungan (fig. S27A). To increase confidence, we also applied Sprime to infer Neanderthal- and Denisovan-derived segments and integrated Neanderthal calls across methods. This yielded 1,605 high-confidence Neanderthal segments spanning 392.51 Mb and 132 Denisovan segments covering 37.17 Mb (fig. S27B). On average, each carried 13.24 Mb of high-confidence Neanderthal sequence and 0.34 Mb of Denisovan sequence (Fig. 6B and fig. S27B). The data support a single Neanderthal admixture pulse and two distinct Denisovan pulses (Fig. 6C).

We next evaluated the phenotypic effects of high-confidence AISs via pathway enrichment and genome-wide association study (GWAS) annotations. Signals are mapped predominantly to metabolic and immune functions. We identified previously unreported Neanderthal signals in immunity-related genes (*CXCR5*, *TLR5*, *PLCL2*, *RAB17*, *USP20*, *TRIML1*, *PIBF1*, *C6*, and *HHLA1*) and metabolism-related genes (*DDC*, *FAH*, *MTRR*, *ASRGL1*, *ACSBG2*, *ALK*, and *CRTC3*). We also replicated many reported Neanderthal associations, including immune genes (*TLR1/10*, *OAS1/2/3*, *CCR1/2/3/5*, *CCRL2*, *ERAP2*, *DOCK2*, and, *LRRC2*), metabolism-related genes (*PASK*, *OSBPL10*, *HDLBP*, and, *TBC1D1*), pigmentation and dermatological traits (*LAMB3* and *TMEM132D*), skin keratinocyte differentiation (*POU2F3*), neurological features (*ADAMTSL3*, *TENM3*, and, *UNC13C*), and reproductive health (*PGR*) (*19, 69-71*). The Denisovan segments included previously reported immune-related genes (*ARHGEF28*, *BANK1*, and *EPHB2*) and additional candidates in immunity (*C1QA/B/C*, *GIMAP4/6/7*, *LECT2*, *DEF6*, *XRCC4*, *PRKCE*), as well as in metabolism (*TCF7L2*, *A1CF*, *GOT2*, *FGF1*, *PPARD*, *HMGCR*, and *SLC38A7*). Among the CAAH subgroups, the Neanderthal segments at *BUD13*, *ZPR1*, and *APOA5* were associated with lipid metabolites. Central Asian-specific Neanderthal sequences with phenotypic relevance included *L3MBTL3* (lipid metabolism), *RAD51B* (RA/asthma), and *ESR1* (bone mineral density) in CAAH-West (*72, 73*); *PPP6R3* (bone mineral density) and *MAD1L1* (psychiatric disorders) in CAAH-Mixed (*74, 75*); *SLCO1B1* (hepatic drug transport) in CAAH-Kyrgyz; and *UBE2L3* (serum lipids and autoimmune disorders

such as Crohn's disease, inflammatory bowel disease, and systemic lupus erythematosus) in CAAH-Dungan *(76)*. Denisovan-derived associations were enriched for height, type 2 diabetes, and Takayasu arteritis; notable variants included *HHAT* rs115453328 in CAAH-West (obsessive–compulsive traits), *RASA3* rs61971965 in CAAH-Kyrgyz (gray matter density), and *IL2RA* rs61839660 in CAAH-Dungan (immune-related traits). Pathway analysis highlighted the shared enrichment in muscle cell differentiation, nervous system development, and muscle tissue growth (fig. S28).

Although archaic adaptive introgression has been reported elsewhere, its impact on Central Asia remains unclear (*4, 66, 67*). We identified a Neanderthal haplotype at *CLOCK* comprising 42 archaic SNPs in CAAH-West, with rs11724094 linked to male-pattern baldness (Fig. 6D). Additional signals in CAAH-West included *CUBN* rs12221306 (metabolite levels) and *WNK2* rs12349094 (educational attainment), each showing distinct temporal frequency trajectories consistent with shifting selection pressures (Fig. 6, E to G). We also detected archaic SNPs at *PXK* rs6445975 (systemic lupus erythematosus), *IL2* rs2069772 (allergic sensitization), and *CDC42BPA* rs1929863 (white blood cell count), with *IL2* rs2069772 exhibiting a marked decline in DAF since ~10 kya (*77*) (Fig. 6H). In CAAH-Kyrgyz, *DGKH* rs73187291 was associated with insulin-like growth factor 1 levels (*42*), and in CAAH-Dungan, *TOR1B* rs12634 (3′ UTR) was linked to prostate cancer. Collectively, these results indicate that archaic introgression is pervasive across Central Asian genomes, materially shaping the genetic architecture of health-related traits and contributing to local adaptation in these populations.

Medical relevance and pharmacogenomic implications

We annotated all variants against ClinVar and identified 116 pathogenic (level 5) variants across 99 genes. Every participant carried at least two ClinVar-listed pathogenic variants (median, 9 alleles; range, 3–16; Fig. 7A). Among the unique pathogenic variants, 56.9% (66/116) were singletons, 16.4% had an MAF < 0.01, and 11.2% had an MAF of 0.01–0.05 (Fig. 7A). Among the 208 pathogenic or likely pathogenic (level 4) variants, 17.8% (37/208) had an MAF > 0.05; three of these had a mean MAF < 0.05 in gnomAD v3.1.2, and only one had an MAF < 0.05 across all gnomAD populations (Fig. 7B and table S8). Most variants (~51.9%) were associated with autosomal recessive disorders, whereas ~20.7% were linked to autosomal dominant disorders (Fig. 7B and fig. S29A). To evaluate clinical utility, we cross-referenced pathogenic and likely pathogenic variants with the American College of Medical Genetics and Genomics (ACMG) Secondary Findings Panel (ACMG-SF v3.2) (*78*). Ninety-two of 166 individuals carried at least one of 10 reportable ACMG variants spanning seven genes (Fig. 7C). Four variants occurred as singletons, in *MUTYH* and *BTD* (autosomal recessive) and in *LDLR* (autosomal dominant), and two appeared as doubletons in *BRCA2* (autosomal dominant) and *ATP7B* (autosomal recessive). Notably, *SCN5A A1673G* (rs1805124, H558R), associated with autosomal dominant ventricular fibrillation and altered cardiac sodium channel function, was present in 65 individuals (~39.2%; Fig. 7B). This variant was most frequent in Karakalpak (~41.2%) and common in African populations (fig. S29B), suggesting population-specific effects or possible misclassification in ACMG and other resources.

We next characterized AFs for pharmacogenomic variants from the ADME gene list and ClinVar drug-response variants, revealing marked differences between Central Asian and other continental populations (Fig. 7D). For warfarin, the derived allele of *VKORC1*-1639A (rs9923231) was frequent (DAF ≥ 73.7%) in Dungan, Hazara, Karakalpak, Kyrgyz, and Uyghur, which is consistent with lower dose requirements in AA homozygotes. *VKORC1* rs7196161 was associated with elevated DAFs (≥ 42.5%) in the Karluk, Tajik, Turkmen, Uzbek, and non-East Asian reference populations (especially South Asians); AA homozygotes may require higher doses. *CYP4F2*3* (rs2108622), which increases the warfarin dose, varied substantially across CAAH populations, as did *CYP2C9*2* (rs1799853), which was enriched in Karluk and Turkmen.

*CYP2C9*2* carriers exhibit reduced warfarin metabolism, increased phenytoin sensitivity, and a greater risk of gastrointestinal bleeding with some non-steroidal anti-inflammatory drugs. Together, these patterns support genotype-guided warfarin dosing in Central Asia. We also observed pronounced AF variation for *ATIC* T675C (rs4673993), with high frequencies (≥ 40%) in Karluk, Tajik, Turkmen, China's Dai, and South Asian references (Fig. 7D and fig. S29C); CC homozygotes presented an enhanced response to methotrexate, underscoring the value of *ATIC* genotyping. Uzbek exhibited a very high DAF at *ADRB2* rs1042713 (77.5%), suggesting reduced salmeterol efficacy in AA homozygotes. Additional AF differences were detected in variants affecting the analgesic response (rs2952768), isoniazid metabolism [*NAT2*5* (rs1801280), *NAT2*6* (rs1799930), *NAT2*12* (rs1208)], interferon response (rs12979860), and salbutamol response (rs6988229).

Loss-of-function variants and human knockouts

Reconstructing the landscape of human homozygous putative loss-of-function (pLoF) alleles can inform therapeutic target discovery and enable assessment of the phenotypic effects of gene loss. We cataloged pLoF variants and human knockouts and identified 3,436 pLoF variants (Fig. 7E): 76.9% frameshift, 12.7% stop-gained, 4.9% splice-donor, and 3.9% splice-acceptor, with the remainder stop-lost. We also detected 3,381 high-confidence protein-truncating variants (PTVs), predominantly heterozygous (96.2%), of which 78.8% were unique to this dataset (Fig. 7E). Gene set enrichment analyses highlighted over-representation of pathways involved in system development, cell motility, and signaling molecules and interactions (Fig. 7F). The enrichment signals were strongest in the Uyghur, including pathways related to cytoskeletal function in muscle cells, head development, sensory system development, and 22q11.2 copy-number variation syndrome (fig. S29D). We further identified 129 homozygous PTVs across 118 genes, eight of which were absent from dbSNP156. The newly observed homozygous PTVs included *CCDC88C* in all CAAH groups; *NOL9* and two *CRK* alleles specific to Uzbek; *ZMIZ2* unique to Tajik; *RELN* restricted to Hazara; and the *BPIFB3* and *TGM2* alleles detected only in Kyrgyz. Notably, a GWAS of 19,629 Europeans linked *RELN* variants to regional brain volumes and cognitive and mental-health traits (*79*).

Discussion

Dataset and fine-scale population structure of CAAH

Central Asia has long been a nexus of trans-Eurasian genetic, cultural, and economic exchange. Despite its pivotal role in shaping present-day Eurasian diversity, the demographic history of contemporary Central Asians and the genetic underpinnings of population-specific traits and diseases remain incompletely resolved. In this study, whole-genome data from 166 individuals across 20 CAGDP populations were analyzed to delineate the demographic and evolutionary processes that shaped the region. Pronounced east–west genetic differentiation in CAAH populations was observed, which is consistent with ancient DNA, historical, and archaeological records (*26, 80, 81*). West Eurasian ancestors, particularly lineages related to Iranian and Anatolian farmers and EEHG, predominated in Tajik, Karluk, Turkmen, and Uzbek. In contrast, Kyrgyz, Uyghur, Karakalpak, and Hazara showed stronger affinities to Altaic groups of East Eurasia and to ancient Mongolian Plateau populations with reduced West Eurasian ancestry. Dungan presented a distinct profile marked by predominant ANEA ancestry and close connections to Sino-Tibetan speakers. An array-based study reporting greater allele sharing between Kyrgyz, Uzbek, and Turkmen populations with diverse East Asian/Siberian groups than with West Eurasians (*32*) partly contrasts with these findings, underscoring the value of high-density genotyping and broad sampling for reconstructing population history. Our large-scale WGS dataset, integrated with publicly available modern and ancient Eurasian genomes, was used to reconstruct the fine-scale and complex deep demographic history of CAAH populations.

Admixture profiles inferred with qpAdm, haplotype-based methods (SOURCEFIND, fastGLOBETROTTER), and fastsimcoal-based evolutionary processes further support stratified ancestries within CAAH: elevated West Eurasian ancestry in Tajik, Karluk, Turkmen, and Uzbek; complex east–west admixture in Karakalpak, Uyghur, and Hazara; enriched East Eurasian ancestry in Kyrgyz; and exceptionally high ANEAs in Dungan. We provide the first coalescence-based evolutionary models of these four genetic lineages, including the Paleolithic divergence between West and East Eurasian populations and the Neolithic divergence within Central Asians. This finding supports the hypothesis that the Tajik-related lineage originated from West Eurasia, Sino-Tibetan-speaking Dungan and Altaic-speaking populations from East Eurasia. These results argue against simplistic historical models and instead indicate prolonged, multidirectional gene flow across Central Asia.

Ancient DNA revealed that WSHG, EEHG, WSH, ANE, Iranian farmer, and Anatolian farmer-related groups were common across Central Asia from the Paleolithic period to the historical period (*25-30*). Combined with the current diversity, this pattern suggests several migration waves and mixing events, including Neolithic farmer expansions, Bronze Age nomadic movements, and later east–west shifts, resulting in more East Eurasian ancestry. The admixture dates (23–34 generations ago) match the Song–Yuan period and the rise of the Mongol Empire, whose westward expansion likely increased East Asian ancestry and changed regional genetic patterns. Different timelines, such as two pulses in Tajiks (~50 and ~9 generations ago), show population-specific paths influenced by unique historical and sociopolitical factors. With the Bronze Age start of east–west admixture (*31*), these estimates are probably conservative lower bounds.

Linguistic and geographic correlations of genetic differentiation

Motivated by recent evidence that genetic variation closely correlates with cultural and geographic features across linguistically diverse Northeast Asian populations, including Uralic and Yeniseian speakers (*82, 83*), we investigate whether similar patterns driven by isolation-by-distance and migration-driven admixture shape Altaic (especially Turkic) populations and their Indo-European neighbors. Substantial heterogeneity was observed within the Turkic groups. Genomic differences follow geographic patterns: Eastern groups (e.g., Salar, Yugur) contrast with western groups (e.g., Tatar, Kumyk, Turkmen). Strong correlations of the first two PCs with longitude and latitude, along with different associations of ancestral components with these axes, indicate that isolation by distance is combined with region-specific admixture. Various hypotheses concerning the origins of Turkic groups include southern Siberia/Mongolia ("steppe hypothesis") (*32*), Northeast China ("farming/Northeast Asia" hypothesis linked to ANEAs and the Xinglongwa/Hongshan cultures) (*84*), and the Altai–Sayan region (*85*). The current findings revealed geographically graded ANA- and ANEA-related ancestry, which decreased from east to west. Admixture modeling identified newly sequenced Turkic individuals as mixtures of Xianbei_IA (~29.7–72.3%) from the ARB and a WSH-derived source, whereas demographic modeling suggested a common ancestor with East Eurasian rather than West Eurasian lineages. Overall, the evidence supports an origin in Northeast Asia with gradual westward movement and cultural diffusion.

Indo-European speakers formed two broad clines, including Eastern Europe and Central Asia/Caucasus. Tajik groups carried elevated EEHG-, Iranian farmer-, and Anatolian farmer-related ancestry, maximizing within West Eurasian Indo-European speakers and reflecting a deep West Eurasian imprint. Prior work reported substantial diversity among Tajiks, including greater allele sharing with specific Bronze Age Xinjiang individuals than with those from Central Asia (*24*). Here, Tajiks are closely aligned with ancient West Eurasian farming and steppe groups and with Turan populations, a pattern that may reflect genuine genetic differentiation among Tajiks and/or reference panel effects. In contrast, Hazara showed increased sharing with Altaic groups

from northern East Asia and southern Siberia, and Dungan aligned more strongly with East Asians, particularly Sino-Tibetan speakers, despite geographic proximity to other Central Asians. Earlier studies placed Hazara with Turkic or Mongolian-related groups (*33, 86*), and historical records described Dungan as a Muslim Hui-derived group migrating from China; recent data confirmed minimal differentiation from southern Ningxia Hui (*87*). These observations underscore the substantial role of long-range migration in shaping the present-day structure.

Signatures of local adaptation

The influence of admixture, selection, and archaic introgression on human health and disease remains a central theme in human genetics. Genome-wide scans for positive selection were performed to infer local adaptation. By integrating AFd_e with ancient DNA time series data, subgroup-specific adaptive trajectories shaped by complex demography, subsistence shifts, and environmental pressures were inferred. The signals following admixture were widespread yet heterogeneous across the four subgroups, with enrichment in metabolic, immune, developmental, and neurological pathways. Convergent selection was observed at *EBF2*, a key regulator of adipogenesis, in CAAH-West and CAAH-Mixed, supporting repeated selection for energy storage and thermoregulation in response to climate change. Signals at *MTAP* in CAAH-Mixed (linked to polyamine metabolism) and *CLDN10* in CAAH-Kyrgyz (related to renal function and intestinal barrier integrity) were also identified. The trajectories of DAFs suggest that shifts in subsistence strategies drove these adaptations.

Immune-related selection involves *TRAF1* and *PHF19*, with disease-associated alleles (e.g., RA risk variants) showing evidence of positive selection in CAAH-West. Given the substantial West Eurasian ancestry, including > 22% EEHG (Fig. 1D), these alleles were likely introduced via admixture, aligning with recent reports of high-frequency RA risk alleles in hunter-gatherers followed by purifying selection over ~15 kya (*88*). In CAAH-Dungan, rs10896794 in *LPXN* (inflammatory bowel disease) was found at a lower-than-expected frequency post-admixture, and DAFs of *GIMAP6* variants plateaued ~3 kya, suggesting recent stabilization of immune-related pressures. The selection of loci related to nervous system function was prominent. Consistent signals at *FILIP1* across the three subgroups suggest regionally convergent pressures on brain-related traits, whereas subgroup-specific signals at *CHRNA3* and *HDAC9* (CAAH-Mixed), *SPON1* (CAAH-Kyrgyz), and *TCF4* (CAAH-Dungan) indicate diverse selection on behavioral, neurological, and cognitive traits. The intersection of AFd_e with East–West HDVs highlighted post-admixture selection at *THOP1* and *VPS41*, emphasizing both localized adaptation and the intricate mosaic of East–West ancestry. Overall, selection across metabolic, immune, and neurodevelopmental pathways reflects various pressures during and after admixture, with a fine-scale structure shaping adaptive landscapes.

Archaic introgression landscape

The extent and functional relevance of Neanderthal and Denisovan introgression in Central Asia were clarified. Neanderthal-derived segment distributions varied across subgroups, which was consistent with distinct demographic and local admixture histories. A total of 104 Neanderthal-inherited segments absent from the 1KGP were identified, underscoring the complexity of introgression and its contribution to regional diversity. CAAH-specific AISs, including a *BCL11A* segment, point to localized pressures potentially affecting hemoglobin levels and susceptibility to sickle cell anemia (*89*). While many phenotypic effects mirrored prior work (notably immune and metabolic loci), additional associations were detected, including Neanderthal-derived variants in immune/metabolic pathways and Denisovan-derived segments linked to neurodevelopmental and psychological traits. Population-specific associations, including *L3MBTL3* and *RAD51B* in CAAH-West (lipid metabolites; immune disorders), *MAD1L1* in CAAH-Mixed (psychiatric traits), *SLCO1B1* in CAAH-Kyrgyz (drug metabolism), and *UBE2L3* in CAAH-Dungan

(autoimmune diseases), suggest distinct health impacts of Neanderthal-derived sequences across subgroups. Pathway analyses revealed shared enrichment for muscle cell differentiation and nervous system/muscle development across Neanderthal- and Denisovan-derived sequences, implying effects on broader physiological processes. Despite widespread AISs, adaptive introgression signals are limited. We identified archaic haplotypes, such as a Neanderthal-derived variant at *CLOCK* in CAAH-West, showing signatures of positive selection. Temporal shifts in DAFs at immune- and metabolism-related loci indicate changing pressures, potentially reflecting subsistence transitions or pathogen landscapes. Many subgroup-specific AISs, particularly in CAAH-Kyrgyz and CAAH-Dungan, lacked strong phenotypic associations, suggesting modest contributions to local adaptation or current detection limits. These results illustrate the value of region-focused genomics for refining the archaic legacy and its evolutionary consequences in historically understudied populations.

Central Asian genomes inform human disease and health

Disease-relevant variation in historically understudied Central Asians has substantially expanded. Numerous rare and population-specific pathogenic variants with clinical relevance have been identified; the high frequency of *SCN5A*-rs1805124 (arrhythmia-associated) illustrates the need for population-tailored interpretation. The catalog of human knockouts was extended with high-confidence LoF alleles and unreported homozygous PTVs that are informative for genotype–phenotype mapping. The frequency spectrum of the pharmacogenomic variants varied markedly across CAAH groups and relative to those of the global panels. Notable differences involved *VKORC1*, *CYP2C9*, and *CYP4F2* (warfarin metabolism), *ATIC* (methotrexate response), *ADRB2* (salmeterol response), and *NAT2*5/*6/*12* (isoniazid response). These data provide a foundation for preemptive screening and for optimizing drug regimens across regional populations.

Materials and Methods

Ethical approval, sample collection, and genome sequencing

This work was developed through collaborations among anthropological, genetic, and linguistic scientists within the CAGDP Consortium, aiming to improve the representation of the genomic diversity of ethnolinguistically diverse populations and to provide comprehensive insights into their population history, biological adaptation, archaic introgression, and medical implications. The study was approved by the Medical Ethics Committee of the National Centre for Biotechnology of Kazakhstan (protocol No. 2 of 10 June 2020) and the Nazarbayev University Institutional Research Ethics Committee (protocol No. 17 of 16 April 2019), and all procedures were conducted in accordance with the Declaration of Helsinki (*90*). The participants were selected based on the criterion that all four grandparents were indigenous to the respective ethnic groups. Informed consent was obtained from all participants. Saliva samples were collected from 166 individuals from five Central Asian countries (Kyrgyzstan, Kazakhstan, Uzbekistan, Tajikistan, and Turkmenistan) and Afghanistan in South Asia via the Oragene DNA Self-Collection Kit (OG-500, DNA Genotek, Canada). The sampled individuals included Kyrgyz people from various regions in Kyrgyzstan (Talas, Issyk-Kul, Naryn, Osh, Batken, and Jalal-Abad), Uzbeks from the Turkistan Region in Kazakhstan, Karakalpaks from the Karakalpakstan Region in Uzbekistan, Karluks from the Khatlon Region in Tajikistan, Uyghurs from the Almaty Region in Kazakhstan, Turkmens from Daşoguz and Mary Regions in Turkmenistan, Dungans from the Jambyl Region in Kazakhstan, Tajiks from several regions in Tajikistan (Sughd, Gorno-Badakhshan, and Khatlon), and Hazaras from various provinces in Afghanistan (Bamyan, Ghazni, Oruzgan, and Ghor). Genomic DNA was extracted via the prepIT-L2P kit (DNA Genotek, Canada) following the manufacturer's protocol, and the DNA concentration was measured via a Qubit 3.0 fluorometer (Thermo Fisher Scientific). Library preparation was performed according to the standard Illumina protocol, with all libraries purified via Agencourt AMPure XP (Beckman

Coulter). The final libraries were quantified via qPCR. WGS was conducted on the Illumina HiSeq 4000 platform, with a preset sequencing depth of 10× for all samples.

Sequencing quality control and variant discovery

Samples were retained if at least 80% of all bases had a quality score of 30 or higher. Sentieon Genomics tools (version 202010.04) were used to process the read mapping of the newly generated genomic data, following the GATK Best Practices guidelines (*91*). The raw sequencing reads were aligned to the human reference genome assemblies GRCh37 and GRCh38 via BWA v0.7.17, respectively (*92*). PCR duplicates were removed with Sentieon Dedup. The aligned sequencing reads from different lanes were then sorted and merged via SAMtools v1.10 (*93*). GVCF files were generated with the HaplotypeCaller module, and joint variant calling was performed via the CombineGVCFs and GenotypeGVCFs modules (*94*). Variant evaluation, including the number of raw or filtered variant counts, the proportion of variants not present in dbSNP, and the ratio of Ti/Tv mutations, was conducted via the VariantEval tool. The identified SNPs were categorized into four groups based on their MAF: (i) singleton variants, (ii) rare variants (MAF < 0.01), (iii) intermediate variants (0.01 ≤ MAF < 0.05), and (iv) common variants (MAF ≥ 0.05).

Dataset merging

The newly generated GRCh37-based dataset was merged with the published HO and 1240K datasets of contemporary and ancient populations worldwide, as included in the Allen Ancient DNA Resource (*37*). Additionally, genotype profiles of Northwest Chinese populations from the pilot study of 10K_CPGDP were incorporated into this study (*7*). Data filtering was performed via PLINK v1.90 (*95*), excluding variants with missing call rates greater than 0.05 (--geno 0.05), MAFs less than 0.05 (--maf 0.05), and exact test *p*-values for Hardy–Weinberg equilibrium less than 10^{-6} (--hwe 10^{-6}). Samples with missing call rates exceeding 0.1 (--mind 0.1) were also removed. Relatedness coefficients between individuals within populations were estimated via KING v.2.3.0 (*96*) and PLINK v.1.90 (*95*). The final merged HO dataset comprised 5,476 individuals and 593,050 SNPs, whereas the merged 1240K dataset included 2,923 individuals and 1,135,264 SNPs. To investigate the fine-scale genetic substructure of Turkic- and Altaic-speaking populations across Eurasia, the merged HO dataset was further integrated with previously published data (*27, 32*). The resulting dataset comprised 5,798 individuals and 81,480 SNPs. The extremely low-density extended HO dataset was used to investigate the population structure of the newly sequenced CAAH populations and their linguistically related groups. The merged low-density HO dataset and the medium-density 1240K dataset were primarily applied for admixture profile modeling and demographic history reconstruction. Furthermore, the GRCh38-based dataset was merged with 929 high-coverage whole-genome sequences from 54 diverse human populations from the Human Genome Diversity Project (HGDP) (*97*) and 317 high-coverage genomes from 20 Oceanian populations (*19*) to create the WGS dataset based on the same pipeline of read mapping and variant calling. Only biallelic autosomal variants were retained for analyses requiring high-density markers, such as inferring effective population sizes, estimating population divergence times, and detecting signatures of natural selection.

Population structure inference

To investigate the genetic affinity and population stratification of CAAH groups within Eurasian and regional contexts, we conducted PCAs using the smartpca program in EIGENSOFT v7.2.1 (*98*), with parameters set to numoutlieriter: 0 and lsqproject: YES. Mlabri and ancient populations were projected onto the first two principal components. Scatter plots were generated using R v4.2.2 and custom scripts. An unsupervised model-based clustering analysis was performed using the maximum likelihood method implemented in ADMIXTURE (v1.3.0) within Eurasian and

regional populations (*99*). The number of ancestral components (K) ranged from 2 to 20, with a 10-fold cross-validation setting (--cv=10). SNPs in strong linkage disequilibrium (LD) were pruned using the "--indep-pairwise 200 25 0.4" command in PLINK v1.90. For each value of K, 100 bootstrap resamplings were conducted with different random seeds. F_{ST} distances between population pairs were calculated using PLINK v1.90, excluding populations with sample sizes less than five. The phylogenetic relationships among ethnolinguistically diverse target populations were further estimated using TreeMix v1.13 (*100*). The frequency spectrum for each population was calculated using PLINK v1.90 and subsequently utilized as input for constructing TreeMix-based topologies. Migration edges, ranging from 0 to 7, were explored using the -m parameter to assess potential gene flow events.

Relationship between genetics and geography among Turkic populations

Geographic coordinates of each sampling site were used to perform Procrustes analyses in the R package vegan (https://github.com/vegandevs/vegan) *(101)*, following the approach of Atkinson et al. *(102)*. The first two PCs from the PCA plot of Turkic groups were rotated and scaled to minimize squared differences between PCs and geographic coordinates after projection onto a sphere. The Procrustes M^2 statistic, representing the sum of squared differences between the two ordinations, was calculated. Statistical significance was assessed using *p*-values derived from 100,000 permutation tests.

F-statistical analysis

Various forms of the f_3- and f_4-statistics were computed using ADMIXTOOLS v.7.0.2 to measure shared genetic drift between two populations since their divergence from a common ancestral population (*103*). Outgroup f_3-statistics of the form f_3(Target, Reference; Mbuti) were performed to estimate overall allele sharing between all target and reference populations. Additionally, admixture f_3-statistics of the form f_3(Ref1, Ref2; Target) were calculated for all pairs of modern or ancient references to test whether the target population was derived from the admixture of populations related to two predefined ancestry surrogates. Tree-like relatedness between four populations was further evaluated using the qpDstat package in ADMIXTOOLS with the f_4 model (f_4Mode: YES). The block jackknife method was employed to assess the statistical significance of the f_4-values. To quantify genetic homogeneity and heterogeneity between two target populations, we conducted symmetrical f_4(Target1, Target2; Ref, Mbuti).

Admixture modeling with qpWave and qpAdm

To determine whether target population pairs (left populations) formed a single clade relative to a set of outgroups (right populations), we conducted rank tests using the qpWave program (version 1520) (*103*). Admixture modeling based on the f_4-statistics was performed using the qpAdm program (version 1520) within the ADMIXTOOLS package. This approach assessed the number of distinct ancestry streams from surrogate populations needed to explain the genetic makeup of the target populations and estimated the corresponding admixture proportions. The parameters "allsnps: YES" and "details: YES" were applied to include all SNPs across the four populations and display the specific entries of the f_4 matrix and Z scores.

Admixture time estimation

Long-distance migrations and subsequent admixtures between populations can lead to an exponential decay of admixture-induced LD. To compute the weighted LD decays and infer the admixture dates, we utilized MALDER v1.0 with additional parameters set to mindis: 0.005 and jackknife: YES (*104*). Multiple modern populations from East and West Eurasia were employed as potential ancestral sources, testing all possible combinations.

Haplogroup classification and clustering analysis
Y-chromosomal haplogroups were classified using Y-LineageTracker 1.3.0 (*105*) and the hGrpr2.py package within HaploGrouper (*106*). For the HaploGrouper-based analysis, two reference files, treeFileNEW_isogg2019.txt and snpFile_b38_isogg2019.txt, were utilized. Maternal haplogroups were classified using HaploGrep2 (*107*) and HaploGrouper, based on PhyloTree17.

Inference of sex-biased admixture
To evaluate potential sex-biased admixture processes in CAAH groups, ADMIXTURE was used to estimate ancestry proportions of surrogate source populations on both autosomes and X chromosomes. Z-scores were calculated to assess differences between autosome- and X chromosome-based admixture proportions for a given ancestry, using the formula $Z = (P_A - P_X)/\sqrt{\sigma_A{}^2 + \sigma_X{}^2}$, where P_A and P_X represent the ancestry proportions on autosomes and X chromosomes, respectively, and σ_A and σ_X are their corresponding standard errors (*108*). A positive Z score indicates a higher ancestry proportion on the autosomes than on the X chromosomes, suggesting a male-driven admixture; conversely, a negative Z score indicates a female-driven admixture. Due to the original use of this formula for assessing differences in qpAdm-based ancestry proportions and the unavailability of X chromosome genotype data for source populations modeled with qpAdm in this study, ADMIXTURE was employed to estimate admixture proportions based on the WGS dataset. Additionally, a Student's t-test was performed (*87*), yielding a *p*-value less than 0.05, which indicates the occurrence of sex-biased admixture in the target population.

Haplotype-based fine-scale population structure reconstruction
Segmented haplotype estimation and fine-scale population structure identification
Stricter filtering criteria of geno: 0.1 and mind: 0.1 were applied to prune genome-wide SNPs. The Segmented HAPlotype Estimation & Imputation Tool (SHAPEIT v2.r904) was then used to estimate haplotypes based on the genetic maps from HapMap phase II b37 (*109, 110*). To identify an optimal starting point for the estimated haplotypes and generate more parsimonious graphs, we performed haplotype phasing with ten burn-in iterations (--burn 10), ten pruning iterations (--prune 10), and 30 main iterations (--main 30) of MCMC (Markov chain Monte Carlo). The fine-scale population structure was subsequently analyzed using fineSTRUCTURE v4.0 based on the phased haplotypes (*111*).

Shared IBD fragment estimation
Refined IBD (refined-ibd.17Jan20.102.jar) was used to detect shared IBD segments between each pair of individuals (*110*). A minimum length of 0.1 cM was set for the identified IBD fragments using the parameter length=0.1. The detected IBD segments were classified into three categories: 1–5 cM, 5–10 cM, and > 10 cM. Segments in the 1–5 cM likely reflect ancient genetic connections dating back 1,500–2,500 years, those in the 5–10 cM correspond to a time interval of approximately 500–1,500 years ago, and segments greater than 10 cM likely indicate recent shared ancestry within the last 500 years. To provide an overview of IBD sharing, we summed the total count and length of IBD segments for each pair of individuals, and the population mean for each block was calculated.

Admixture events inferred from fastGLOBETROTTER
ChromoPainterv2 was used to paint each chromosome of the recipient populations as a series of haplotype chunks and to estimate the number of ancestry chunks inherited from donor populations (*111*). Initially, ChromoPainterv2 was run to paint each phased haploid of the first ten recipient individuals, using ten iterations of expectation-maximization with the "-in" switch and "-iM"

emission parameters to estimate the chunk size and switch error rate. A Perl script was subsequently applied to estimate the switch (-n) and emission (-M) rates across individuals. The estimated parameters were then used to rerun ChromoPainterv2 to construct the copy vectors, modeling both recipient and donor individuals as a patchwork of donor haplotypes. The "chunk length" output files obtained across all chromosomes were summed. FastGLOBETROTTER was then employed to identify, date, and describe admixture events in target populations, using the chunk length output and painting samples, with the options "prop.ind: 1" and "null.ind: 1" (*112*). The significance of the inferred admixture dates was estimated using "bootstrap.date.ind: 1" and "bootstrap.num: 100" options. SourcefindV2 was also used to infer admixture models based on haplotype sharing identified by ChromoPainter.

Demographic history inference

For each newly sequenced population, SMC++ was used to infer the population size history (*113*). Input files were prepared based on the genotypes of all called biallelic SNPs and included all individuals from each population. The two samples with the highest sequencing coverage in each population were specified as the distinguished lineages. SMC++ was run with a mutation rate of 1.25×10^{-8} per base pair per generation and a constant generation time of 29 years. To infer the population divergence time, MSMC2 (multiple sequentially Markovian coalescent) was employed (*114*). The two samples with the highest sequencing coverage were selected from each target population for MSMC2 analysis. The French and Han Chinese populations from the HGDP were chosen as representatives of European and East Asian ancestries, respectively. MSMC2 analyses were conducted following standard recommendations (*113*), utilizing mappability masks for the human reference genome GRCh38 generated by the SNPable pipeline. MSMC2 was run three times independently to estimate coalescence rates within population1/population2 and across populations. Absolute time estimates were calculated using the same mutation rate and generation time as those adopted in SMC++. The relative cross-coalescence rate (rCCR) was calculated, and the divergence time for each population pair was inferred at the point where the rCCR reached 0.5.

Complex demographic model reconstruction

To investigate the complex demographic history of Central Asian populations, demographic models were reconstructed via fastsimcoal v2.7 (*115*). The French and Han populations from the HGDP were selected as proxies for European and East Asian ancestral lineages, respectively. From each of the four CAAH subgroups, one representative population, Tajik, Karakalpak, Kyrgyz, and Dungan, was included to capture the regional genetic diversity of four representative genetic lineages. SNPs were filtered according to the following criteria: (i) missing in at least one sample; (ii) deviation from Hardy–Weinberg equilibrium ($p < 10^{-4}$) in at least one population; (iii) located within a CpG island; (iv) located in a coding region; and (v) absence of ancestral state information. Pairwise site frequency spectra (SFS) were generated for these populations, and we evaluated multiple demographic models incorporating plausible divergence and gene flow scenarios: (i) Tajik shares a common ancestor with French, whereas Karakalpak, Kyrgyz, and Dungan share a common ancestor with Han, with Karakalpak and Kyrgyz forming a lineage that diverged earlier than Dungan and Han with or without instant western Eurasian gene flow events (modelA/B); (ii) Tajik, Karakalpak, and Kyrgyz share a common ancestor with French, with the common ancestor of Karakalpak and Kyrgyz diverging first with or without eastern Eurasian gene flow event (modelC/D); and (iii) Karakalpak and Kyrgyz originate from an East Eurasian lineage with the common ancestor of Han and Dungan, but later received instantaneous gene flow from either French or Tajik (modelE/F/G). Model parameters were estimated by maximizing the composite likelihood of the observed SFS under each model, with multiple independent runs performed to ensure convergence and robustness. Parameter estimates were scaled using a

mutation rate of 1.25×10^{-8} per site per generation and a generation time of 29 years. The best-fitting model was identified based on likelihood values, providing quantitative estimates of the timing and magnitude of divergence and admixture events that have shaped the genetic landscape of Central Asia.

Medical relevance and biological adaptation

Variant annotation

Variant annotation of the GRCh37-based dataset was performed using Ensembl VEP, with a focus on variants annotated as pathogenic or likely pathogenic (*116*). These pathogenic variants were reviewed against the ACMG-SF v3.2 gene panel following the criteria outlined by the ACMG (*117*). SnpEff v4.3, with default parameters, was employed for LoF predictions (*118*), defining a variant as a PTV if it was annotated as "frameshift", "transcript ablation", "splice-acceptor", "splice-donor", or "stop-gained". To generate a list of high-confidence PTVs, we restricted variants to high-confidence regions as determined by Genome in a Bottle (ftp://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/giab/data/AshkenazimTrio/analysis/NIST_v4.1_SmallVariantDraftBenchmark_12182019/GRCh37/HG002_GRCh37_1_22_v4.1_draft_benchmark.be).

Estimation of frequency deviations and HDVs

To screen for signals of natural selection, CAAH populations were divided into four subgroups based on haplotype-based clustering patterns. CAAH-West primarily encompassed populations such as Tajik, Karluk, Turkmen, and Uzbek, which exhibited a greater proportion of West Eurasian ancestry. CAAH-Mixed mainly included Uyghur, Hazara, and Karakalpak, showing a balanced mixture of West and East Eurasian ancestries. CAAH-Kyrgyz consisted of Kyrgyz people, characterized by a relatively higher proportion of East Eurasian ancestry, while CAAH-Dungan predominantly comprised Dungan people, which displayed a strong genetic affinity with East Asian populations. The AFd_e of genome-wide variations within CAAH groups was estimated based on the absolute difference between AF_{obs} and AF_{exp}. The AF_{exp} was calculated as the admixture proportion weighted by the frequencies in ancestral populations. Weights were assigned based on global admixture proportions estimated through ADMIXTURE analysis. A window-based approach, as described in previous studies (*119*), was employed to identify variants exhibiting substantial AFd_e, which serve as genomic signatures of local adaptation.

Signals of natural selection were further identified by screening for HDVs between East and West Eurasians with substantial frequency deviations. CAAH was treated as a simple composite of east–west Eurasian ancestral populations to facilitate HDV identification. The Han Chinese and French populations from the HGDP were utilized as representative ancestral populations of East and West Eurasia, respectively. Pairwise F_{ST} values of overlapping SNPs between these ancestral populations were calculated to identify HDVs, with SNPs having F_{ST} values greater than 0.2 classified as HDVs. The frequency deviation was measured by calculating AFd_e ($|AF_{obs} - AF_{exp}|$), where $AF_{exp} = f_{East} \times \alpha + f_{West} \times (1-\alpha)$ (*38, 39*). Here, f_{East} and f_{West} represent the AFs of a specific SNP in the East and West Eurasian ancestral populations, respectively, while α denotes the admixture proportion of East Eurasian ancestry in CAAH, and $(1-\alpha)$ represents the proportion of West Eurasian ancestry. HDVs exhibiting an AFd_e greater than 0.1 in CAAH groups suggest potential adaptive significance. Functional annotations of candidate variants were performed using the GWAS Catalog (v.2022-09-14), and pathway enrichment was assessed with Metascape (*120*). The frequency trajectory of the detected natural selection signal was visualized via the AGES browser (https://reich-ages.rc.hms.harvard.edu/#/). To identify evidence of natural selection, X-statistic values derived from ancient DNA time series data were applied (*48*).

Archaic introgression estimation

The IBDmix (v.1.0.1) method was initially employed to detect AISs likely derived from Neanderthals (*68*). Owing to the limited sensitivity of IBDmix in detecting Denisovan-derived AISs, Sprime v.07Dec18.5e2 was applied exclusively to identify AISs in CAAH that are likely derived from Denisovans (*121*). Additionally, Sprime was utilized to detect Neanderthal-derived AISs. The newly generated data were merged with 1KGP WGS data for IBDmix analysis. All multi-allelic variants and InDels from the archaic genome and the merged dataset were removed, with a focus on autosomal variants. Archaic ancestry detection was then conducted within each CAAH subgroup to mitigate the impact of population structure. To minimize misclassification due to incomplete lineage sorting, only AISs with a logarithm of the odds ratio for the linkage score above 4 and a length greater than 50 kb were included in downstream analyses. To maximize the detection of Neanderthal-like sequences, a conservative approach was used for filtering the callset. After the initial identification of Neanderthal- and Denisovan-derived sequences via IBDmix, the Neanderthal-related callset was refined by masking regions identified as Denisovan-like segments in Africans that were also classified as Neanderthal-derived sequences in any population.

For the Sprime analysis, the four previously defined CAAH subgroups were designated target populations, with the YRI population (Yoruba in Ibadan, Nigeria, n = 108) serving as the outgroup. A score threshold of 150,000 and default parameters were employed to optimize performance and accuracy. To ascertain the origin of AISs and the proportion of putative archaic alleles matching reference sequences, we calculated the match rate for each AIS, which was defined as the ratio of matched sites to the total number of compared sites. A contour plot was generated via the kde2d function from the MASS package in R to visualize distinct waves of archaic introgression. Segments with clear Neanderthal or Denisovan origins were extracted for further exploration of the biological functions of the identified AISs. Only segments with at least 30 putatively introgressed alleles comparable to the Vindija Neanderthal genome or at least 30 comparable to the Denisovan genome were considered. Neanderthal-derived segments were defined as those with a match rate exceeding 0.6 with the Vindija Neanderthal genome and below 0.3 with the Altai Denisovan genome. Conversely, Denisovan-derived segments were characterized by a match rate above 0.3 with the Altai Denisovan genome and below 0.3 with the Vindija Neanderthal genome. To minimize misclassification due to incomplete lineage sorting, only AISs with a logarithm of the odds ratio for the linkage score above 4 and a length greater than 50 kb were included in downstream analyses (*122*). Finally, the Neanderthal-inherited AISs identified via both methods were intersected via bedtools to generate a high-confidence dataset for adaptive introgression signal identification (*123*). To obtain high-confidence signals of adaptive introgression, we also employed haplotype-based iHS to examine the genomic signatures of natural selection. Estimates of iHS scores were conducted via Selscan v1.2.0 (*124*), with normalization performed via norm v1.2.1 with the parameters "--bp-win --winsize 100000 --qbins 10". SNPs with normalized absolute iHS values exceeding two were selected as candidates for selective sweeps. The focus was restricted to archaic variants detected by both frequency- and haplotype-based selection signals.

References

1. R. A. Gibbs, The Human Genome Project changed everything. *Nat Rev Genet* **21**, 575-576 (2020).
2. D. Gurdasani, T. Carstensen, S. Fatumo, G. Chen, C. S. Franklin, J. Prado-Martinez, H. Bouman, F. Abascal, M. Haber, I. Tachmazidou, I. Mathieson, K. Ekoru, M. K. DeGorter, R. N. Nsubuga, C. Finan, E. Wheeler, L. Chen, D. N. Cooper, S. Schiffels, Y. Chen, G. R. S. Ritchie, M. O. Pollard, M. D. Fortune, A. J. Mentzer, E. Garrison, A. Bergstrom, K. Hatzikotoulas, A. Adeyemo, A. Doumatey, H. Elding, L. V. Wain, G. Ehret, P. L. Auer, C. L. Kooperberg, A. P. Reiner, N. Franceschini, D. Maher, S. B. Montgomery, C. Kadie,

C. Widmer, Y. Xue, J. Seeley, G. Asiki, A. Kamali, E. H. Young, C. Pomilla, N. Soranzo, E. Zeggini, F. Pirie, A. P. Morris, D. Heckerman, C. Tyler-Smith, A. A. Motala, C. Rotimi, P. Kaleebu, I. Barroso, M. S. Sandhu, Uganda Genome Resource Enables Insights into Population History and Genomic Discovery in Africa. *Cell* **179**, 984-1002 e1036 (2019).

3. D. Wu, J. Dou, X. Chai, C. Bellis, A. Wilm, C. C. Shih, W. W. J. Soon, N. Bertin, C. B. Lin, C. C. Khor, M. DeGiorgio, S. Cheng, L. Bao, N. Karnani, W. Y. K. Hwang, S. Davila, P. Tan, A. Shabbir, A. Moh, E. K. Tan, J. N. Foo, L. L. Goh, K. P. Leong, R. S. Y. Foo, C. S. P. Lam, A. M. Richards, C. Y. Cheng, T. Aung, T. Y. Wong, H. H. Ng, S. K. Consortium, J. Liu, C. Wang, Large-Scale Whole-Genome Sequencing of Three Diverse Asian Populations in Singapore. *Cell* **179**, 736-749 e715 (2019).
4. K. C. GenomeAsia, The GenomeAsia 100K Project enables genetic discoveries across Asia. *Nature* **576**, 106-111 (2019).
5. S. Mallick, H. Li, M. Lipson, I. Mathieson, M. Gymrek, F. Racimo, M. Zhao, N. Chennagiri, S. Nordenfelt, A. Tandon, P. Skoglund, I. Lazaridis, S. Sankararaman, Q. Fu, N. Rohland, G. Renaud, Y. Erlich, T. Willems, C. Gallo, J. P. Spence, Y. S. Song, G. Poletti, F. Balloux, G. van Driem, P. de Knijff, I. G. Romero, A. R. Jha, D. M. Behar, C. M. Bravi, C. Capelli, T. Hervig, A. Moreno-Estrada, O. L. Posukh, E. Balanovska, O. Balanovsky, S. Karachanak-Yankova, H. Sahakyan, D. Toncheva, L. Yepiskoposyan, C. Tyler-Smith, Y. Xue, M. S. Abdullah, A. Ruiz-Linares, C. M. Beall, A. Di Rienzo, C. Jeong, E. B. Starikovskaya, E. Metspalu, J. Parik, R. Villems, B. M. Henn, U. Hodoglugil, R. Mahley, A. Sajantila, G. Stamatoyannopoulos, J. T. Wee, R. Khusainova, E. Khusnutdinova, S. Litvinov, G. Ayodo, D. Comas, M. F. Hammer, T. Kivisild, W. Klitz, C. A. Winkler, D. Labuda, M. Bamshad, L. B. Jorde, S. A. Tishkoff, W. S. Watkins, M. Metspalu, S. Dryomov, R. Sukernik, L. Singh, K. Thangaraj, S. Paabo, J. Kelso, N. Patterson, D. Reich, The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. *Nature* **538**, 201-206 (2016).
6. S. Nurk, S. Koren, A. Rhie, M. Rautiainen, A. V. Bzikadze, A. Mikheenko, M. R. Vollger, N. Altemose, L. Uralsky, A. Gershman, S. Aganezov, S. J. Hoyt, M. Diekhans, G. A. Logsdon, M. Alonge, S. E. Antonarakis, M. Borchers, G. G. Bouffard, S. Y. Brooks, G. V. Caldas, N. C. Chen, H. Cheng, C. S. Chin, W. Chow, L. G. de Lima, P. C. Dishuck, R. Durbin, T. Dvorkina, I. T. Fiddes, G. Formenti, R. S. Fulton, A. Fungtammasan, E. Garrison, P. G. S. Grady, T. A. Graves-Lindsay, I. M. Hall, N. F. Hansen, G. A. Hartley, M. Haukness, K. Howe, M. W. Hunkapiller, C. Jain, M. Jain, E. D. Jarvis, P. Kerpedjiev, M. Kirsche, M. Kolmogorov, J. Korlach, M. Kremitzki, H. Li, V. V. Maduro, T. Marschall, A. M. McCartney, J. McDaniel, D. E. Miller, J. C. Mullikin, E. W. Myers, N. D. Olson, B. Paten, P. Peluso, P. A. Pevzner, D. Porubsky, T. Potapova, E. I. Rogaev, J. A. Rosenfeld, S. L. Salzberg, V. A. Schneider, F. J. Sedlazeck, K. Shafin, C. J. Shew, A. Shumate, Y. Sims, A. F. A. Smit, D. C. Soto, I. Sovic, J. M. Storer, A. Streets, B. A. Sullivan, F. Thibaud-Nissen, J. Torrance, J. Wagner, B. P. Walenz, A. Wenger, J. M. D. Wood, C. Xiao, S. M. Yan, A. C. Young, S. Zarate, U. Surti, R. C. McCoy, M. Y. Dennis, I. A. Alexandrov, J. L. Gerton, R. J. O'Neill, W. Timp, J. M. Zook, M. C. Schatz, E. E. Eichler, K. H. Miga, A. M. Phillippy, The complete sequence of a human genome. *Science* **376**, 44-53 (2022).
7. G. He, H. Yao, S. Duan, L. Luo, Q. Sun, R. Tang, J. Chen, Z. Wang, Y. Sun, X. Li, L. Hu, L. Yun, J. Yang, J. Yan, S. Nie, Y. Zhu, K. C. Consortium, C. C. Wang, B. Liu, L. Hu, C. Liu, M. Wang, 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).

8. X. Li, M. Wang, H. Su, S. Duan, Y. Sun, H. Chen, Z. Wang, Q. Sun, Q. Yang, J. Chen, T. Yang, Y. Huang, J. Zhong, X. Jiang, J. Ma, S. Chen, Y. Liu, L. Luo, X. Lin, S. Nie, C. Liu, L. Hu, G. He, Evolutionary history and biological adaptation of Han Chinese people on the Mongolian Plateau. *hLife* **2**, 296-313 (2024).
9. L. Luo, M. Wang, Y. Liu, J. Li, F. Bu, H. Yuan, R. Tang, C. Liu, G. He, Sequencing and characterizing human mitochondrial genomes in the biobank-based genomic research paradigm. *Sci China Life Sci* **68**, 1610-1625 (2025).
10. S. Fatumo, T. Chikowore, A. Choudhury, M. Ayub, A. R. Martin, K. Kuchenbaecker, A roadmap to increase diversity in genomic studies. *Nat Med* **28**, 243-250 (2022).
11. G. He, M. Wang, L. Luo, Q. Sun, H. Yuan, H. Lv, Y. Feng, X. Liu, J. Cheng, F. Bu, M. Zhabagin, H. Yuan, C. Liu, S. Xu, Population genomics of Central Asian peoples unveil ancient Trans-Eurasian genetic admixture and cultural exchanges. *hLife* **2**, 554-562 (2024).
12. M. Wang, S. Duan, X. Li, J. Yang, H. Yuan, C. Liu, G. He, Genome-driven Chinese precision medicine: Biobank-scale genomic research as a new paradigm. *The Innovation Life* **3**, (2025).
13. D. Ju, D. Hui, D. A. Hammond, A. Wonkam, S. A. Tishkoff, Importance of Including Non-European Populations in Large Human Genetic Studies to Enhance Precision Medicine. *Annu Rev Biomed Data Sci* **5**, 321-339 (2022).
14. G. Sirugo, S. M. Williams, S. A. Tishkoff, The Missing Diversity in Human Genetic Studies. *Cell* **177**, 26-31 (2019).
15. U. K. B. W.-G. S. Consortium, Whole-genome sequencing of 490,640 UK Biobank participants. *Nature*, (2025).
16. D. Taliun, D. N. Harris, M. D. Kessler, J. Carlson, Z. A. Szpiech, R. Torres, S. A. G. Taliun, A. Corvelo, S. M. Gogarten, H. M. Kang, A. N. Pitsillides, J. LeFaive, S. B. Lee, X. Tian, B. L. Browning, S. Das, A. K. Emde, W. E. Clarke, D. P. Loesch, A. C. Shetty, T. W. Blackwell, A. V. Smith, Q. Wong, X. Liu, M. P. Conomos, D. M. Bobo, F. Aguet, C. Albert, A. Alonso, K. G. Ardlie, D. E. Arking, S. Aslibekyan, P. L. Auer, J. Barnard, R. G. Barr, L. Barwick, L. C. Becker, R. L. Beer, E. J. Benjamin, L. F. Bielak, J. Blangero, M. Boehnke, D. W. Bowden, J. A. Brody, E. G. Burchard, B. E. Cade, J. F. Casella, B. Chalazan, D. I. Chasman, Y. I. Chen, M. H. Cho, S. H. Choi, M. K. Chung, C. B. Clish, A. Correa, J. E. Curran, B. Custer, D. Darbar, M. Daya, M. de Andrade, D. L. DeMeo, S. K. Dutcher, P. T. Ellinor, L. S. Emery, C. Eng, D. Fatkin, T. Fingerlin, L. Forer, M. Fornage, N. Franceschini, C. Fuchsberger, S. M. Fullerton, S. Germer, M. T. Gladwin, D. J. Gottlieb, X. Guo, M. E. Hall, J. He, N. L. Heard-Costa, S. R. Heckbert, M. R. Irvin, J. M. Johnsen, A. D. Johnson, R. Kaplan, S. L. R. Kardia, T. Kelly, S. Kelly, E. E. Kenny, D. P. Kiel, R. Klemmer, B. A. Konkle, C. Kooperberg, A. Kottgen, L. A. Lange, J. Lasky-Su, D. Levy, X. Lin, K. H. Lin, C. Liu, R. J. F. Loos, L. Garman, R. Gerszten, S. A. Lubitz, K. L. Lunetta, A. C. Y. Mak, A. Manichaikul, A. K. Manning, R. A. Mathias, D. D. McManus, S. T. McGarvey, J. B. Meigs, D. A. Meyers, J. L. Mikulla, M. A. Minear, B. D. Mitchell, S. Mohanty, M. E. Montasser, C. Montgomery, A. C. Morrison, J. M. Murabito, A. Natale, P. Natarajan, S. C. Nelson, K. E. North, J. R. O'Connell, N. D. Palmer, N. Pankratz, G. M. Peloso, P. A. Peyser, J. Pleiness, W. S. Post, B. M. Psaty, D. C. Rao, S. Redline, A. P. Reiner, D. Roden, J. I. Rotter, I. Ruczinski, C. Sarnowski, S. Schoenherr, D. A. Schwartz, J. S. Seo, S. Seshadri, V. A. Sheehan, W. H. Sheu, M. B. Shoemaker, N. L. Smith, J. A. Smith, N. Sotoodehnia, A. M. Stilp, W. Tang, K. D. Taylor, M. Telen, T. A. Thornton, R. P. Tracy, D. J. Van Den Berg, R. S. Vasan, K. A. Viaud-Martinez, S. Vrieze, D. E. Weeks, B. S. Weir, S. T. Weiss, L. C. Weng, C. J. Willer, Y. Zhang, X. Zhao, D. K. Arnett, A. E. Ashley-Koch, K. C. Barnes, E. Boerwinkle, S. Gabriel, R. Gibbs, K. M. Rice, S. S. Rich, E. K. Silverman, P. Qasba, W. Gan, N. T.-O. f. P. M.

Consortium, G. J. Papanicolaou, D. A. Nickerson, S. R. Browning, M. C. Zody, S. Zollner, J. G. Wilson, L. A. Cupples, C. C. Laurie, C. E. Jaquish, R. D. Hernandez, T. D. O'Connor, G. R. Abecasis, Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature* **590**, 290-299 (2021).

17. A. Choudhury, S. Aron, L. R. Botigue, D. Sengupta, G. Botha, T. Bensellak, G. Wells, J. Kumuthini, D. Shriner, Y. J. Fakim, A. W. Ghoorah, E. Dareng, T. Odia, O. Falola, E. Adebiyi, S. Hazelhurst, G. Mazandu, O. A. Nyangiri, M. Mbiyavanga, A. Benkahla, S. K. Kassim, N. Mulder, S. N. Adebamowo, E. R. Chimusa, D. Muzny, G. Metcalf, R. A. Gibbs, G. E. N. R. G. Trypano, C. Rotimi, M. Ramsay, H. A. Consortium, A. A. Adeyemo, Z. Lombard, N. A. Hanchard, High-depth African genomes inform human migration and health. *Nature* **586**, 741-748 (2020).
18. M. Sohail, M. J. Palma-Martinez, A. Y. Chong, C. D. Quinto-Cortes, C. Barberena-Jonas, S. G. Medina-Munoz, A. Ragsdale, G. Delgado-Sanchez, L. P. Cruz-Hervert, L. Ferreyra-Reyes, E. Ferreira-Guerrero, N. Mongua-Rodriguez, S. Canizales-Quintero, A. Jimenez-Kaufmann, H. Moreno-Macias, C. A. Aguilar-Salinas, K. Auckland, A. Cortes, V. Acuna-Alonzo, C. R. Gignoux, G. L. Wojcik, A. G. Ioannidis, S. L. Fernandez-Valverde, A. V. S. Hill, M. T. Tusie-Luna, A. J. Mentzer, J. Novembre, L. Garcia-Garcia, A. Moreno-Estrada, Mexican Biobank advances population and medical genomics of diverse ancestries. *Nature* **622**, 775-783 (2023).
19. J. Choin, J. Mendoza-Revilla, L. R. Arauna, S. Cuadros-Espinoza, O. Cassar, M. Larena, A. M. Ko, C. Harmant, R. Laurent, P. Verdu, G. Laval, A. Boland, R. Olaso, J. F. Deleuze, F. Valentin, Y. C. Ko, M. Jakobsson, A. Gessain, L. Excoffier, M. Stoneking, E. Patin, L. Quintana-Murci, Genomic insights into population history and biological adaptation in Oceania. *Nature* **592**, 583-589 (2021).
20. A. Nagai, M. Hirata, Y. Kamatani, K. Muto, K. Matsuda, Y. Kiyohara, T. Ninomiya, A. Tamakoshi, Z. Yamagata, T. Mushiroda, Y. Murakami, K. Yuji, Y. Furukawa, H. Zembutsu, T. Tanaka, Y. Ohnishi, Y. Nakamura, G. BioBank Japan Cooperative Hospital, M. Kubo, Overview of the BioBank Japan Project: Study design and profile. *J Epidemiol* **27**, S2-S8 (2017).
21. R. G. Walters, I. Y. Millwood, K. Lin, D. Schmidt Valle, P. McDonnell, A. Hacker, D. Avery, A. Edris, H. Fry, N. Cai, W. W. Kretzschmar, M. A. Ansari, P. A. Lyons, R. Collins, P. Donnelly, M. Hill, R. Peto, H. Shen, X. Jin, C. Nie, X. Xu, Y. Guo, C. Yu, J. Lv, R. J. Clarke, L. Li, Z. Chen, G. China Kadoorie Biobank Collaborative, Genotyping and population characteristics of the China Kadoorie Biobank. *Cell Genom* **3**, 100361 (2023).
22. P. K. Cong, W. Y. Bai, J. C. Li, M. Y. Yang, S. Khederzadeh, S. R. Gai, N. Li, Y. H. Liu, S. H. Yu, W. W. Zhao, J. Q. Liu, Y. Sun, X. W. Zhu, P. P. Zhao, J. W. Xia, P. L. Guan, Y. Qian, J. G. Tao, L. Xu, G. Tian, P. Y. Wang, S. Y. Xie, M. C. Qiu, K. Q. Liu, B. S. Tang, H. F. Zheng, Genomic analyses of 10,376 individuals in the Westlake BioBank for Chinese (WBBC) pilot project. *Nat Commun* **13**, 2939 (2022).
23. Y. Cao, L. Li, M. Xu, Z. Feng, X. Sun, J. Lu, Y. Xu, P. Du, T. Wang, R. Hu, Z. Ye, L. Shi, X. Tang, L. Yan, Z. Gao, G. Chen, Y. Zhang, L. Chen, G. Ning, Y. Bi, W. Wang, M. A. P. C. China, The ChinaMAP analytics of deep whole genome sequences in 10,588 individuals. *Cell Res* **30**, 717-731 (2020).
24. S. S. Dai, X. Sulaiman, J. Isakova, W. F. Xu, N. T. Abdulloevich, M. E. Afanasevna, K. B. Ibrohimovich, X. Chen, W. K. Yang, M. S. Wang, Q. K. Shen, X. Y. Yang, Y. G. Yao, A. A. Aldashev, A. Saidov, W. Chen, L. F. Cheng, M. S. Peng, Y. P. Zhang, The Genetic Echo of the Tarim Mummies in Modern Central Asians. *Mol Biol Evol* **39**, (2022).
25. G. A. Gnecchi-Ruscone, E. Khussainova, N. Kahbatkyzy, L. Musralina, M. A. Spyrou, R. A. Bianco, R. Radzeviciute, N. F. G. Martins, C. Freund, O. Iksan, A. Garshin, Z.

Zhaniyazov, B. Bekmanov, E. Kitov, Z. Samashev, A. Beisenov, N. Berezina, Y. Berezin, A. Z. Biro, S. Evinger, A. Bissembaev, G. Akhatov, A. Mamedov, A. Onggaruly, D. Voyakin, A. Chotbayev, Y. Kariyev, A. Buzhilova, L. Djansugurova, C. Jeong, J. Krause, Ancient genomic time transect from the Central Asian Steppe unravels the history of the Scythians. *Sci Adv* **7**, (2021).

26. V. M. Narasimhan, N. Patterson, P. Moorjani, N. Rohland, R. Bernardos, S. Mallick, I. Lazaridis, N. Nakatsuka, I. Olalde, M. Lipson, A. M. Kim, L. M. Olivieri, A. Coppa, M. Vidale, J. Mallory, V. Moiseyev, E. Kitov, J. Monge, N. Adamski, N. Alex, N. Broomandkhoshbacht, F. Candilio, K. Callan, O. Cheronet, B. J. Culleton, M. Ferry, D. Fernandes, S. Freilich, B. Gamarra, D. Gaudio, M. Hajdinjak, E. Harney, T. K. Harper, D. Keating, A. M. Lawson, M. Mah, K. Mandl, M. Michel, M. Novak, J. Oppenheimer, N. Rai, K. Sirak, V. Slon, K. Stewardson, F. Zalzala, Z. Zhang, G. Akhatov, A. N. Bagashev, A. Bagnera, B. Baitanayev, J. Bendezu-Sarmiento, A. A. Bissembaev, G. L. Bonora, T. T. Chargynov, T. Chikisheva, P. K. Dashkovskiy, A. Derevianko, M. Dobes, K. Douka, N. Dubova, M. N. Duisengali, D. Enshin, A. Epimakhov, A. V. Fribus, D. Fuller, A. Goryachev, A. Gromov, S. P. Grushin, B. Hanks, M. Judd, E. Kazizov, A. Khokhlov, A. P. Krygin, E. Kupriyanova, P. Kuznetsov, D. Luiselli, F. Maksudov, A. M. Mamedov, T. B. Mamirov, C. Meiklejohn, D. C. Merrett, R. Micheli, O. Mochalov, S. Mustafokulov, A. Nayak, D. Pettener, R. Potts, D. Razhev, M. Rykun, S. Sarno, T. M. Savenkova, K. Sikhymbaeva, S. M. Slepchenko, O. A. Soltobaev, N. Stepanova, S. Svyatko, K. Tabaldiev, M. Teschler-Nicola, A. A. Tishkin, V. V. Tkachev, S. Vasilyev, P. Veleminsky, D. Voyakin, A. Yermolayeva, M. Zahir, V. S. Zubkov, A. Zubova, V. S. Shinde, C. Lalueza-Fox, M. Meyer, D. Anthony, N. Boivin, K. Thangaraj, D. J. Kennett, M. Frachetti, R. Pinhasi, D. Reich, The formation of human populations in South and Central Asia. *Science* **365**, (2019).

27. C. Jeong, O. Balanovsky, E. Lukianova, N. Kahbatkyzy, P. Flegontov, V. Zaporozhchenko, A. Immel, C. C. Wang, O. Ixan, E. Khussainova, B. Bekmanov, V. Zaibert, M. Lavryashina, E. Pocheshkhova, Y. Yusupov, A. Agdzhoyan, S. Koshel, A. Bukin, P. Nymadawa, S. Turdikulova, D. Dalimova, M. Churnosov, R. Skhalyakho, D. Daragan, Y. Bogunov, A. Bogunova, A. Shtrunov, N. Dubova, M. Zhabagin, L. Yepiskoposyan, V. Churakov, N. Pislegin, L. Damba, L. Saroyants, K. Dibirova, L. Atramentova, O. Utevska, E. Idrisov, E. Kamenshchikova, I. Evseeva, M. Metspalu, A. K. Outram, M. Robbeets, L. Djansugurova, E. Balanovska, S. Schiffels, W. Haak, D. Reich, J. Krause, The genetic history of admixture across inner Eurasia. *Nat Ecol Evol* **3**, 966-976 (2019).

28. P. de Barros Damgaard, R. Martiniano, J. Kamm, J. V. Moreno-Mayar, G. Kroonen, M. Peyrot, G. Barjamovic, S. Rasmussen, C. Zacho, N. Baimukhanov, V. Zaibert, V. Merz, A. Biddanda, I. Merz, V. Loman, V. Evdokimov, E. Usmanova, B. Hemphill, A. Seguin-Orlando, F. E. Yediay, I. Ullah, K. G. Sjogren, K. H. Iversen, J. Choin, C. de la Fuente, M. Ilardo, H. Schroeder, V. Moiseyev, A. Gromov, A. Polyakov, S. Omura, S. Y. Senyurt, H. Ahmad, C. McKenzie, A. Margaryan, A. Hameed, A. Samad, N. Gul, M. H. Khokhar, O. I. Goriunova, V. I. Bazaliiskii, J. Novembre, A. W. Weber, L. Orlando, M. E. Allentoft, R. Nielsen, K. Kristiansen, M. Sikora, A. K. Outram, R. Durbin, E. Willerslev, The first horse herders and the impact of early Bronze Age steppe expansions into Asia. *Science* **360**, (2018).

29. P. B. Damgaard, N. Marchi, S. Rasmussen, M. Peyrot, G. Renaud, T. Korneliussen, J. V. Moreno-Mayar, M. W. Pedersen, A. Goldberg, E. Usmanova, N. Baimukhanov, V. Loman, L. Hedeager, A. G. Pedersen, K. Nielsen, G. Afanasiev, K. Akmatov, A. Aldashev, A. Alpaslan, G. Baimbetov, V. I. Bazaliiskii, A. Beisenov, B. Boldbaatar, B. Boldgiv, C. Dorzhu, S. Ellingvag, D. Erdenebaatar, R. Dajani, E. Dmitriev, V.

Evdokimov, K. M. Frei, A. Gromov, A. Goryachev, H. Hakonarson, T. Hegay, Z. Khachatryan, R. Khaskhanov, E. Kitov, A. Kolbina, T. Kubatbek, A. Kukushkin, I. Kukushkin, N. Lau, A. Margaryan, I. Merkyte, I. V. Mertz, V. K. Mertz, E. Mijiddorj, V. Moiyesev, G. Mukhtarova, B. Nurmukhanbetov, Z. Orozbekova, I. Panyushkina, K. Pieta, V. Smrcka, I. Shevnina, A. Logvin, K. G. Sjogren, T. Stolcova, A. M. Taravella, K. Tashbaeva, A. Tkachev, T. Tulegenov, D. Voyakin, L. Yepiskoposyan, S. Undrakhbold, V. Varfolomeev, A. Weber, M. A. Wilson Sayres, N. Kradin, M. E. Allentoft, L. Orlando, R. Nielsen, M. Sikora, E. Heyer, K. Kristiansen, E. Willerslev, 137 ancient human genomes from across the Eurasian steppes. *Nature* **557**, 369-374 (2018).

30. V. Kumar, E. A. Bennett, D. Zhao, Y. Liang, Y. Tang, M. Ren, Q. Dai, X. Feng, P. Cao, R. Yang, F. Liu, W. Ping, M. Zhang, M. Ding, M. A. Yang, B. Amridin, H. Muttalib, J. Wang, Q. Fu, Genetic Continuity of Bronze Age Ancestry with Increased Steppe-Related Ancestry in Late Iron Age Uzbekistan. *Mol Biol Evol* **38**, 4908-4917 (2021).
31. C. Jeong, K. Wang, S. Wilkin, W. T. T. Taylor, B. K. Miller, J. H. Bemmann, R. Stahl, C. Chiovelli, F. Knolle, S. Ulziibayar, D. Khatanbaatar, D. Erdenebaatar, U. Erdenebat, A. Ochir, G. Ankhsanaa, C. Vanchigdash, B. Ochir, C. Munkhbayar, D. Tumen, A. Kovalev, N. Kradin, B. A. Bazarov, D. A. Miyagashev, P. B. Konovalov, E. Zhambaltarova, A. V. Miller, W. Haak, S. Schiffels, J. Krause, N. Boivin, M. Erdene, J. Hendy, C. Warinner, A Dynamic 6,000-Year Genetic History of Eurasia's Eastern Steppe. *Cell* **183**, 890-904 e829 (2020).
32. B. Yunusbayev, M. Metspalu, E. Metspalu, A. Valeev, S. Litvinov, R. Valiev, V. Akhmetova, E. Balanovska, O. Balanovsky, S. Turdikulova, D. Dalimova, P. Nymadawa, A. Bahmanimehr, H. Sahakyan, K. Tambets, S. Fedorova, N. Barashkov, I. Khidiyatova, E. Mihailov, R. Khusainova, L. Damba, M. Derenko, B. Malyarchuk, L. Osipova, M. Voevoda, L. Yepiskoposyan, T. Kivisild, E. Khusnutdinova, R. Villems, The genetic legacy of the expansion of Turkic-speaking nomads across Eurasia. *PLoS Genet* **11**, e1005068 (2015).
33. B. Martinez-Cruz, R. Vitalis, L. Segurel, F. Austerlitz, M. Georges, S. Thery, L. Quintana-Murci, T. Hegay, A. Aldashev, F. Nasyrova, E. Heyer, In the heartland of Eurasia: the multilocus genetic landscape of Central Asian populations. *Eur J Hum Genet* **19**, 216-223 (2011).
34. G. Hellenthal, G. B. J. Busby, G. Band, J. F. Wilson, C. Capelli, D. Falush, S. Myers, A genetic atlas of human admixture history. *Science* **343**, 747-751 (2014).
35. Q. Feng, Y. Lu, X. Ni, K. Yuan, Y. Yang, X. Yang, C. Liu, H. Lou, Z. Ning, Y. Wang, D. Lu, C. Zhang, Y. Zhou, M. Shi, L. Tian, X. Wang, X. Zhang, J. Li, A. Khan, Y. Guan, K. Tang, S. Wang, S. Xu, Genetic History of Xinjiang's Uyghurs Suggests Bronze Age Multiple-Way Contacts in Eurasia. *Mol Biol Evol* **34**, 2572-2582 (2017).
36. P. Guarino-Vignon, N. Marchi, J. Bendezu-Sarmiento, E. Heyer, C. Bon, Genetic continuity of Indo-Iranian speakers since the Iron Age in southern Central Asia. *Sci Rep* **12**, 733 (2022).
37. S. Mallick, A. Micco, M. Mah, H. Ringbauer, I. Lazaridis, I. Olalde, N. Patterson, D. Reich, The Allen Ancient DNA Resource (AADR) a curated compendium of ancient human genomes. *Sci Data* **11**, 182 (2024).
38. Y. Pan, C. Zhang, Y. Lu, Z. Ning, D. Lu, Y. Gao, X. Zhao, Y. Yang, Y. Guan, D. Mamatyusupu, S. Xu, Genomic diversity and post-admixture adaptation in the Uyghurs. *Natl Sci Rev* **9**, nwab124 (2022).
39. Z. Ning, X. Tan, Y. Yuan, K. Huang, Y. Pan, L. Tian, Y. Lu, X. Wang, R. Qi, D. Lu, Y. Yang, Y. Guan, D. Mamatyusupu, S. Xu, Expression profiles of east-west highly differentiated genes in Uyghur genomes. *Natl Sci Rev* **10**, nwad077 (2023).

40. E. Jorgenson, N. Makki, L. Shen, D. C. Chen, C. Tian, W. L. Eckalbar, D. Hinds, N. Ahituv, A. Avins, A genome-wide association study identifies four novel susceptibility loci underlying inguinal hernia. *Nat Commun* **6**, 10130 (2015).
41. P. van der Harst, N. Verweij, Identification of 64 Novel Genetic Loci Provides an Expanded View on the Genetic Architecture of Coronary Artery Disease. *Circ Res* **122**, 433-443 (2018).
42. N. Sinnott-Armstrong, Y. Tanigawa, D. Amar, N. Mars, C. Benner, M. Aguirre, G. R. Venkataraman, M. Wainberg, H. M. Ollila, T. Kiiskinen, A. S. Havulinna, J. P. Pirruccello, J. Qian, A. Shcherbina, FinnGen, F. Rodriguez, T. L. Assimes, V. Agarwala, R. Tibshirani, T. Hastie, S. Ripatti, J. K. Pritchard, M. J. Daly, M. A. Rivas, Genetics of 35 blood and urine biomarkers in the UK Biobank. *Nat Genet* **53**, 185-194 (2021).
43. S. Sakaue, M. Kanai, Y. Tanigawa, J. Karjalainen, M. Kurki, S. Koshiba, A. Narita, T. Konuma, K. Yamamoto, M. Akiyama, K. Ishigaki, A. Suzuki, K. Suzuki, W. Obara, K. Yamaji, K. Takahashi, S. Asai, Y. Takahashi, T. Suzuki, N. Shinozaki, H. Yamaguchi, S. Minami, S. Murayama, K. Yoshimori, S. Nagayama, D. Obata, M. Higashiyama, A. Masumoto, Y. Koretsune, FinnGen, K. Ito, C. Terao, T. Yamauchi, I. Komuro, T. Kadowaki, G. Tamiya, M. Yamamoto, Y. Nakamura, M. Kubo, Y. Murakami, K. Yamamoto, Y. Kamatani, A. Palotie, M. A. Rivas, M. J. Daly, K. Matsuda, Y. Okada, A cross-population atlas of genetic associations for 220 human phenotypes. *Nat Genet* **53**, 1415-1424 (2021).
44. G. Kichaev, G. Bhatia, P. R. Loh, S. Gazal, K. Burch, M. K. Freund, A. Schoech, B. Pasaniuc, A. L. Price, Leveraging Polygenic Functional Enrichment to Improve GWAS Power. *Am J Hum Genet* **104**, 65-75 (2019).
45. Y. Han, Q. Jia, P. S. Jahani, B. P. Hurrell, C. Pan, P. Huang, J. Gukasyan, N. C. Woodward, E. Eskin, F. D. Gilliland, O. Akbari, J. A. Hartiala, H. Allayee, Genome-wide analysis highlights contribution of immune system pathways to the genetic architecture of asthma. *Nat Commun* **11**, 1776 (2020).
46. E. Ha, S. C. Bae, K. Kim, Large-scale meta-analysis across East Asian and European populations updated genetic architecture and variant-driven biology of rheumatoid arthritis, identifying 11 novel susceptibility loci. *Ann Rheum Dis* **80**, 558-565 (2021).
47. A. Johansson, M. Rask-Andersen, T. Karlsson, W. E. Ek, Genome-wide association analysis of 350 000 Caucasians from the UK Biobank identifies novel loci for asthma, hay fever and eczema. *Hum Mol Genet* **28**, 4022-4041 (2019).
48. A. Akbari, A. R. Barton, S. Gazal, Z. Li, M. Kariminejad, A. Perry, Y. Zeng, A. Mittnik, N. Patterson, M. Mah, X. Zhou, A. L. Price, E. S. Lander, R. Pinhasi, N. Rohland, S. Mallick, D. Reich, Pervasive findings of directional selection realize the promise of ancient DNA to elucidate human adaptation. *bioRxiv*, (2024).
49. L. Jostins, S. Ripke, R. K. Weersma, R. H. Duerr, D. P. McGovern, K. Y. Hui, J. C. Lee, L. P. Schumm, Y. Sharma, C. A. Anderson, J. Essers, M. Mitrovic, K. Ning, I. Cleynen, E. Theatre, S. L. Spain, S. Raychaudhuri, P. Goyette, Z. Wei, C. Abraham, J. P. Achkar, T. Ahmad, L. Amininejad, A. N. Ananthakrishnan, V. Andersen, J. M. Andrews, L. Baidoo, T. Balschun, P. A. Bampton, A. Bitton, G. Boucher, S. Brand, C. Buning, A. Cohain, S. Cichon, M. D'Amato, D. De Jong, K. L. Devaney, M. Dubinsky, C. Edwards, D. Ellinghaus, L. R. Ferguson, D. Franchimont, K. Fransen, R. Gearry, M. Georges, C. Gieger, J. Glas, T. Haritunians, A. Hart, C. Hawkey, M. Hedl, X. Hu, T. H. Karlsen, L. Kupcinskas, S. Kugathasan, A. Latiano, D. Laukens, I. C. Lawrance, C. W. Lees, E. Louis, G. Mahy, J. Mansfield, A. R. Morgan, C. Mowat, W. Newman, O. Palmieri, C. Y. Ponsioen, U. Potocnik, N. J. Prescott, M. Regueiro, J. I. Rotter, R. K. Russell, J. D. Sanderson, M. Sans, J. Satsangi, S. Schreiber, L. A. Simms, J. Sventoraityte, S. R. Targan, K. D. Taylor, M. Tremelling, H. W. Verspaget, M. De Vos, C. Wijmenga, D. C. Wilson, J.

Winkelmann, R. J. Xavier, S. Zeissig, B. Zhang, C. K. Zhang, H. Zhao, I. B. D. G. C. International, M. S. Silverberg, V. Annese, H. Hakonarson, S. R. Brant, G. Radford-Smith, C. G. Mathew, J. D. Rioux, E. E. Schadt, M. J. Daly, A. Franke, M. Parkes, S. Vermeire, J. C. Barrett, J. H. Cho, Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature* **491**, 119-124 (2012).

50. X. Han, J. S. Ong, J. An, A. W. Hewitt, P. Gharahkhani, S. MacGregor, Using Mendelian randomization to evaluate the causal relationship between serum C-reactive protein levels and age-related macular degeneration. *Eur J Epidemiol* **35**, 139-146 (2020).
51. A. A. Shadrin, T. Kaufmann, D. van der Meer, C. E. Palmer, C. Makowski, R. Loughnan, T. L. Jernigan, T. M. Seibert, D. J. Hagler, O. B. Smeland, E. Motazedi, Y. Chu, A. Lin, W. Cheng, G. Hindley, W. K. Thompson, C. C. Fan, D. Holland, L. T. Westlye, O. Frei, O. A. Andreassen, A. M. Dale, Vertex-wise multivariate genome-wide association study identifies 780 unique genetic loci associated with cortical morphology. *Neuroimage* **244**, 118603 (2021).
52. P. Sakornsakolpat, M. McCormack, P. Bakke, A. Gulsvik, B. J. Make, J. D. Crapo, M. H. Cho, E. K. Silverman, Genome-Wide Association Analysis of Single-Breath Dl(CO). *Am J Respir Cell Mol Biol* **60**, 523-531 (2019).
53. M. Liu, Y. Jiang, R. Wedow, Y. Li, D. M. Brazel, F. Chen, G. Datta, J. Davila-Velderrain, D. McGuire, C. Tian, X. Zhan, T. andMe Research, H. A.-I. Psychiatry, H. Choquet, A. R. Docherty, J. D. Faul, J. R. Foerster, L. G. Fritsche, M. E. Gabrielsen, S. D. Gordon, J. Haessler, J. J. Hottenga, H. Huang, S. K. Jang, P. R. Jansen, Y. Ling, R. Magi, N. Matoba, G. McMahon, A. Mulas, V. Orru, T. Palviainen, A. Pandit, G. W. Reginsson, A. H. Skogholt, J. A. Smith, A. E. Taylor, C. Turman, G. Willemsen, H. Young, K. A. Young, G. J. M. Zajac, W. Zhao, W. Zhou, G. Bjornsdottir, J. D. Boardman, M. Boehnke, D. I. Boomsma, C. Chen, F. Cucca, G. E. Davies, C. B. Eaton, M. A. Ehringer, T. Esko, E. Fiorillo, N. A. Gillespie, D. F. Gudbjartsson, T. Haller, K. M. Harris, A. C. Heath, J. K. Hewitt, I. B. Hickie, J. E. Hokanson, C. J. Hopfer, D. J. Hunter, W. G. Iacono, E. O. Johnson, Y. Kamatani, S. L. R. Kardia, M. C. Keller, M. Kellis, C. Kooperberg, P. Kraft, K. S. Krauter, M. Laakso, P. A. Lind, A. Loukola, S. M. Lutz, P. A. F. Madden, N. G. Martin, M. McGue, M. B. McQueen, S. E. Medland, A. Metspalu, K. L. Mohlke, J. B. Nielsen, Y. Okada, U. Peters, T. J. C. Polderman, D. Posthuma, A. P. Reiner, J. P. Rice, E. Rimm, R. J. Rose, V. Runarsdottir, M. C. Stallings, A. Stancakova, H. Stefansson, K. K. Thai, H. A. Tindle, T. Tyrfingsson, T. L. Wall, D. R. Weir, C. Weisner, J. B. Whitfield, B. S. Winsvold, J. Yin, L. Zuccolo, L. J. Bierut, K. Hveem, J. J. Lee, M. R. Munafo, N. L. Saccone, C. J. Willer, M. C. Cornelis, S. P. David, D. A. Hinds, E. Jorgenson, J. Kaprio, J. A. Stitzel, K. Stefansson, T. E. Thorgeirsson, G. Abecasis, D. J. Liu, S. Vrieze, Association studies of up to 1.2 million individuals yield new insights into the genetic etiology of tobacco and alcohol use. *Nat Genet* **51**, 237-244 (2019).
54. K. L. Grasby, N. Jahanshad, J. N. Painter, L. Colodro-Conde, J. Bralten, D. P. Hibar, P. A. Lind, F. Pizzagalli, C. R. K. Ching, M. A. B. McMahon, N. Shatokhina, L. C. P. Zsembik, S. I. Thomopoulos, A. H. Zhu, L. T. Strike, I. Agartz, S. Alhusaini, M. A. A. Almeida, D. Alnaes, I. K. Amlien, M. Andersson, T. Ard, N. J. Armstrong, A. Ashley-Koch, J. R. Atkins, M. Bernard, R. M. Brouwer, E. E. L. Buimer, R. Bulow, C. Burger, D. M. Cannon, M. Chakravarty, Q. Chen, J. W. Cheung, B. Couvy-Duchesne, A. M. Dale, S. Dalvie, T. K. de Araujo, G. I. de Zubicaray, S. M. C. de Zwarte, A. den Braber, N. T. Doan, K. Dohm, S. Ehrlich, H. R. Engelbrecht, S. Erk, C. C. Fan, I. O. Fedko, S. F. Foley, J. M. Ford, M. Fukunaga, M. E. Garrett, T. Ge, S. Giddaluru, A. L. Goldman, M. J. Green, N. A. Groenewold, D. Grotegerd, T. P. Gurholt, B. A. Gutman, N. K. Hansell, M. A. Harris, M. B. Harrison, C. C. Haswell, M. Hauser, S. Herms, D. J. Heslenfeld, N. F. Ho, D. Hoehn, P. Hoffmann, L. Holleran, M. Hoogman, J. J. Hottenga, M. Ikeda, D. Janowitz,

I. E. Jansen, T. Jia, C. Jockwitz, R. Kanai, S. Karama, D. Kasperaviciute, T. Kaufmann, S. Kelly, M. Kikuchi, M. Klein, M. Knapp, A. R. Knodt, B. Kramer, M. Lam, T. M. Lancaster, P. H. Lee, T. A. Lett, L. B. Lewis, I. Lopes-Cendes, M. Luciano, F. Macciardi, A. F. Marquand, S. R. Mathias, T. R. Melzer, Y. Milaneschi, N. Mirza-Schreiber, J. C. V. Moreira, T. W. Muhleisen, B. Muller-Myhsok, P. Najt, S. Nakahara, K. Nho, L. M. Olde Loohuis, D. P. Orfanos, J. F. Pearson, T. L. Pitcher, B. Putz, Y. Quide, A. Ragothaman, F. M. Rashid, W. R. Reay, R. Redlich, C. S. Reinbold, J. Repple, G. Richard, B. C. Riedel, S. L. Risacher, C. S. Rocha, N. R. Mota, L. Salminen, A. Saremi, A. J. Saykin, F. Schlag, L. Schmaal, P. R. Schofield, R. Secolin, C. Y. Shapland, L. Shen, J. Shin, E. Shumskaya, I. E. Sonderby, E. Sprooten, K. E. Tansey, A. Teumer, A. Thalamuthu, D. Tordesillas-Gutierrez, J. A. Turner, A. Uhlmann, C. L. Vallerga, D. van der Meer, M. M. J. van Donkelaar, L. van Eijk, T. G. M. van Erp, N. E. M. van Haren, D. van Rooij, M. J. van Tol, J. H. Veldink, E. Verhoef, E. Walton, M. Wang, Y. Wang, J. M. Wardlaw, W. Wen, L. T. Westlye, C. D. Whelan, S. H. Witt, K. Wittfeld, C. Wolf, T. Wolfers, J. Q. Wu, C. L. Yasuda, D. Zaremba, Z. Zhang, M. P. Zwiers, E. Artiges, A. A. Assareh, R. Ayesa-Arriola, A. Belger, C. L. Brandt, G. G. Brown, S. Cichon, J. E. Curran, G. E. Davies, F. Degenhardt, M. F. Dennis, B. Dietsche, S. Djurovic, C. P. Doherty, R. Espiritu, D. Garijo, Y. Gil, P. A. Gowland, R. C. Green, A. N. Hausler, W. Heindel, B. C. Ho, W. U. Hoffmann, F. Holsboer, G. Homuth, N. Hosten, C. R. Jack, Jr., M. Jang, A. Jansen, N. A. Kimbrel, K. Kolskar, S. Koops, A. Krug, K. O. Lim, J. J. Luykx, D. H. Mathalon, K. A. Mather, V. S. Mattay, S. Matthews, J. Mayoral Van Son, S. C. McEwen, I. Melle, D. W. Morris, B. A. Mueller, M. Nauck, J. E. Nordvik, M. M. Nothen, D. S. O'Leary, N. Opel, M. P. Martinot, G. B. Pike, A. Preda, E. B. Quinlan, P. E. Rasser, V. Ratnakar, S. Reppermund, V. M. Steen, P. A. Tooney, F. R. Torres, D. J. Veltman, J. T. Voyvodic, R. Whelan, T. White, H. Yamamori, H. H. H. Adams, J. C. Bis, S. Debette, C. Decarli, M. Fornage, V. Gudnason, E. Hofer, M. A. Ikram, L. Launer, W. T. Longstreth, O. L. Lopez, B. Mazoyer, T. H. Mosley, G. V. Roshchupkin, C. L. Satizabal, R. Schmidt, S. Seshadri, Q. Yang, I. Alzheimer's Disease Neuroimaging, C. Consortium, E. Consortium, I. Consortium, S. Y. S. Consortium, I. Parkinson's Progression Markers, M. K. M. Alvim, D. Ames, T. J. Anderson, O. A. Andreassen, A. Arias-Vasquez, M. E. Bastin, B. T. Baune, J. C. Beckham, J. Blangero, D. I. Boomsma, H. Brodaty, H. G. Brunner, R. L. Buckner, J. K. Buitelaar, J. R. Bustillo, W. Cahn, M. J. Cairns, V. Calhoun, V. J. Carr, X. Caseras, S. Caspers, G. L. Cavalleri, F. Cendes, A. Corvin, B. Crespo-Facorro, J. C. Dalrymple-Alford, U. Dannlowski, E. J. C. de Geus, I. J. Deary, N. Delanty, C. Depondt, S. Desrivieres, G. Donohoe, T. Espeseth, G. Fernandez, S. E. Fisher, H. Flor, A. J. Forstner, C. Francks, B. Franke, D. C. Glahn, R. L. Gollub, H. J. Grabe, O. Gruber, A. K. Haberg, A. R. Hariri, C. A. Hartman, R. Hashimoto, A. Heinz, F. A. Henskens, M. H. J. Hillegers, P. J. Hoekstra, A. J. Holmes, L. E. Hong, W. D. Hopkins, H. E. Hulshoff Pol, T. L. Jernigan, E. G. Jonsson, R. S. Kahn, M. A. Kennedy, T. T. J. Kircher, P. Kochunov, J. B. J. Kwok, S. Le Hellard, C. M. Loughland, N. G. Martin, J. L. Martinot, C. McDonald, K. L. McMahon, A. Meyer-Lindenberg, P. T. Michie, R. A. Morey, B. Mowry, L. Nyberg, J. Oosterlaan, R. A. Ophoff, C. Pantelis, T. Paus, Z. Pausova, B. Penninx, T. J. C. Polderman, D. Posthuma, M. Rietschel, J. L. Roffman, L. M. Rowland, P. S. Sachdev, P. G. Samann, U. Schall, G. Schumann, R. J. Scott, K. Sim, S. M. Sisodiya, J. W. Smoller, I. E. Sommer, B. St Pourcain, D. J. Stein, A. W. Toga, J. N. Trollor, N. J. A. Van der Wee, D. van 't Ent, H. Volzke, H. Walter, B. Weber, D. R. Weinberger, M. J. Wright, J. Zhou, J. L. Stein, P. M. Thompson, S. E. Medland, g. Enhancing NeuroImaging Genetics through Meta-Analysis Consortium -Genetics working, The genetic architecture of the human cerebral cortex. *Science* **367**, eaay6690 (2020).

55. D. van der Meer, O. Frei, T. Kaufmann, A. A. Shadrin, A. Devor, O. B. Smeland, W. K. Thompson, C. C. Fan, D. Holland, L. T. Westlye, O. A. Andreassen, A. M. Dale, Understanding the genetic determinants of the brain with MOSTest. *Nat Commun* **11**, 3512 (2020).
56. C. C. Fan, R. Loughnan, C. Makowski, D. Pecheva, C. H. Chen, D. J. Hagler, Jr., W. K. Thompson, N. Parker, D. van der Meer, O. Frei, O. A. Andreassen, A. M. Dale, Multivariate genome-wide association study on tissue-sensitive diffusion metrics highlights pathways that shape the human brain. *Nat Commun* **13**, 2423 (2022).
57. C. X. Yap, J. Sidorenko, Y. Wu, K. E. Kemper, J. Yang, N. R. Wray, M. R. Robinson, P. M. Visscher, Dissection of genetic variation and evidence for pleiotropy in male pattern baldness. *Nat Commun* **9**, 5407 (2018).
58. J. Liu, Y. Zhou, S. Liu, X. Song, X. Z. Yang, Y. Fan, W. Chen, Z. C. Akdemir, Z. Yan, Y. Zuo, R. Du, Z. Liu, B. Yuan, S. Zhao, G. Liu, Y. Chen, Y. Zhao, M. Lin, Q. Zhu, Y. Niu, P. Liu, S. Ikegawa, Y. Q. Song, J. E. Posey, G. Qiu, D. Study, F. Zhang, Z. Wu, J. R. Lupski, N. Wu, The coexistence of copy number variations (CNVs) and single nucleotide polymorphisms (SNPs) at a locus can result in distorted calculations of the significance in associating SNPs to disease. *Hum Genet* **137**, 553-567 (2018).
59. N. Jahanshad, P. Rajagopalan, X. Hua, D. P. Hibar, T. M. Nir, A. W. Toga, C. R. Jack, Jr., A. J. Saykin, R. C. Green, M. W. Weiner, S. E. Medland, G. W. Montgomery, N. K. Hansell, K. L. McMahon, G. I. de Zubicaray, N. G. Martin, M. J. Wright, P. M. Thompson, I. Alzheimer's Disease Neuroimaging, Genome-wide scan of healthy human connectome discovers SPON1 gene variant influencing dementia severity. *Proc Natl Acad Sci U S A* **110**, 4768-4773 (2013).
60. M. Nagel, P. R. Jansen, S. Stringer, K. Watanabe, C. A. de Leeuw, J. Bryois, J. E. Savage, A. R. Hammerschlag, N. G. Skene, A. B. Munoz-Manchado, T. andMe Research, T. White, H. Tiemeier, S. Linnarsson, J. Hjerling-Leffler, T. J. C. Polderman, P. F. Sullivan, S. van der Sluis, D. Posthuma, Meta-analysis of genome-wide association studies for neuroticism in 449,484 individuals identifies novel genetic loci and pathways. *Nat Genet* **50**, 920-927 (2018).
61. R. Karlsson Linner, P. Biroli, E. Kong, S. F. W. Meddens, R. Wedow, M. A. Fontana, M. Lebreton, S. P. Tino, A. Abdellaoui, A. R. Hammerschlag, M. G. Nivard, A. Okbay, C. A. Rietveld, P. N. Timshel, M. Trzaskowski, R. Vlaming, C. L. Zund, Y. Bao, L. Buzdugan, A. H. Caplin, C. Y. Chen, P. Eibich, P. Fontanillas, J. R. Gonzalez, P. K. Joshi, V. Karhunen, A. Kleinman, R. Z. Levin, C. M. Lill, G. A. Meddens, G. Muntane, S. Sanchez-Roige, F. J. V. Rooij, E. Taskesen, Y. Wu, F. Zhang, T. and Me Research, Q. C. e, C. International Cannabis, C. Social Science Genetic Association, A. Auton, J. D. Boardman, D. W. Clark, A. Conlin, C. C. Dolan, U. Fischbacher, P. J. F. Groenen, K. M. Harris, G. Hasler, A. Hofman, M. A. Ikram, S. Jain, R. Karlsson, R. C. Kessler, M. Kooyman, J. MacKillop, M. Mannikko, C. Morcillo-Suarez, M. B. McQueen, K. M. Schmidt, M. C. Smart, M. Sutter, A. R. Thurik, A. G. Uitterlinden, J. White, H. Wit, J. Yang, L. Bertram, D. I. Boomsma, T. Esko, E. Fehr, D. A. Hinds, M. Johannesson, M. Kumari, D. Laibson, P. K. E. Magnusson, M. N. Meyer, A. Navarro, A. A. Palmer, T. H. Pers, D. Posthuma, D. Schunk, M. B. Stein, R. Svento, H. Tiemeier, P. Timmers, P. Turley, R. J. Ursano, G. G. Wagner, J. F. Wilson, J. Gratten, J. J. Lee, D. Cesarini, D. J. Benjamin, P. D. Koellinger, J. P. Beauchamp, Genome-wide association analyses of risk tolerance and risky behaviors in over 1 million individuals identify hundreds of loci and shared genetic influences. *Nat Genet* **51**, 245-257 (2019).
62. K. J. A. Johnston, M. J. Adams, B. I. Nicholl, J. Ward, R. J. Strawbridge, A. Ferguson, A. M. McIntosh, M. E. S. Bailey, D. J. Smith, Genome-wide association study of multisite chronic pain in UK Biobank. *PLoS Genet* **15**, e1008164 (2019).

63. K. Watanabe, P. R. Jansen, J. E. Savage, P. Nandakumar, X. Wang, T. andMe Research, D. A. Hinds, J. Gelernter, D. F. Levey, R. Polimanti, M. B. Stein, E. J. W. Van Someren, A. B. Smit, D. Posthuma, Genome-wide meta-analysis of insomnia prioritizes genes associated with metabolic and psychiatric pathways. *Nat Genet* **54**, 1125-1132 (2022).
64. S. P. Hagenaars, W. D. Hill, S. E. Harris, S. J. Ritchie, G. Davies, D. C. Liewald, C. R. Gale, D. J. Porteous, I. J. Deary, R. E. Marioni, Genetic prediction of male pattern baldness. *PLoS Genet* **13**, e1006594 (2017).
65. C. Tziotzios, C. Petridis, N. Dand, C. Ainali, J. R. Saklatvala, V. Pullabhatla, A. Onoufriadis, R. Pramanik, D. Baudry, S. H. Lee, K. Wood, L. Liu, S. Seegobin, G. A. Michelotti, S. M. Lwin, E. A. A. Christou, C. J. Curtis, E. de Rinaldis, A. Saxena, S. Holmes, M. Harries, I. Palamaras, F. Cunningham, G. Parkins, M. Kaur, P. Farrant, A. McDonagh, A. Messenger, J. Jones, V. Jolliffe, I. Ali, M. Ardern-Jones, C. Mitchell, N. Burrows, R. Atkar, C. Banfield, A. Alexandroff, C. Champagne, H. L. Cooper, S. Vano-Galvan, A. M. Molina-Ruiz, N. O. Perez, G. K. Patel, A. Macbeth, M. Page, A. Bryden, M. Mowbray, S. Wahie, K. Armstrong, N. Cooke, M. Goodfield, I. Man, D. de Berker, G. Dunnill, A. Takwale, A. Rao, T. W. Siah, R. Sinclair, M. S. Wade, N. C. Dlova, J. Setterfield, F. Lewis, K. Bhargava, N. Kirkpatrick, X. Estivill, C. M. Stefanato, C. Flohr, T. Spector, F. M. Watt, C. H. Smith, J. N. Barker, D. A. Fenton, M. A. Simpson, J. A. McGrath, Genome-wide association study in frontal fibrosing alopecia identifies four susceptibility loci including HLA-B*07:02. *Nat Commun* **10**, 1150 (2019).
66. S. Peyregne, V. Slon, J. Kelso, More than a decade of genetic research on the Denisovans. *Nat Rev Genet* **25**, 83-103 (2024).
67. H. Zeberg, M. Jakobsson, S. Paabo, The genetic changes that shaped Neandertals, Denisovans, and modern humans. *Cell* **187**, 1047-1058 (2024).
68. L. Chen, A. B. Wolf, W. Fu, L. Li, J. M. Akey, Identifying and Interpreting Apparent Neanderthal Ancestry in African Individuals. *Cell* **180**, 677-687 e616 (2020).
69. P. F. Reilly, A. Tjahjadi, S. L. Miller, J. M. Akey, S. Tucci, The contribution of Neanderthal introgression to modern human traits. *Curr Biol* **32**, R970-R983 (2022).
70. S. Rong, C. R. Neil, A. Welch, C. Duan, S. Maguire, I. C. Meremikwu, M. Meyerson, B. J. Evans, W. G. Fairbrother, Large-scale functional screen identifies genetic variants with splicing effects in modern and archaic humans. *Proc Natl Acad Sci U S A* **120**, e2218308120 (2023).
71. F. Racimo, D. Marnetto, E. Huerta-Sanchez, Signatures of Archaic Adaptive Introgression in Present-Day Human Populations. *Mol Biol Evol* **34**, 296-317 (2017).
72. L. A. Lotta, P. Gulati, F. R. Day, F. Payne, H. Ongen, M. van de Bunt, K. J. Gaulton, J. D. Eicher, S. J. Sharp, J. Luan, E. De Lucia Rolfe, I. D. Stewart, E. Wheeler, S. M. Willems, C. Adams, H. Yaghootkar, E. P.-I. Consortium, F. C. Cambridge, N. G. Forouhi, K. T. Khaw, A. D. Johnson, R. K. Semple, T. Frayling, J. R. Perry, E. Dermitzakis, M. I. McCarthy, I. Barroso, N. J. Wareham, D. B. Savage, C. Langenberg, S. O'Rahilly, R. A. Scott, Integrative genomic analysis implicates limited peripheral adipose storage capacity in the pathogenesis of human insulin resistance. *Nat Genet* **49**, 17-26 (2017).
73. K. Ishigaki, S. Sakaue, C. Terao, Y. Luo, K. Sonehara, K. Yamaguchi, T. Amariuta, C. L. Too, V. A. Laufer, I. C. Scott, S. Viatte, M. Takahashi, K. Ohmura, A. Murasawa, M. Hashimoto, H. Ito, M. Hammoudeh, S. A. Emadi, B. K. Masri, H. Halabi, H. Badsha, I. W. Uthman, X. Wu, L. Lin, T. Li, D. Plant, A. Barton, G. Orozco, S. M. M. Verstappen, J. Bowes, A. J. MacGregor, S. Honda, M. Koido, K. Tomizuka, Y. Kamatani, H. Tanaka, E. Tanaka, A. Suzuki, Y. Maeda, K. Yamamoto, S. Miyawaki, G. Xie, J. Zhang, C. I. Amos, E. Keystone, G. Wolbink, I. van der Horst-Bruinsma, J. Cui, K. P. Liao, R. J. Carroll, H. S. Lee, S. Y. Bang, K. A. Siminovitch, N. de Vries, L. Alfredsson, S. Rantapaa-Dahlqvist, E. W. Karlson, S. C. Bae, R. P. Kimberly, J. C. Edberg, X. Mariette, T. Huizinga, P.

Dieude, M. Schneider, M. Kerick, J. C. Denny, P. BioBank Japan, K. Matsuda, K. Matsuo, T. Mimori, F. Matsuda, K. Fujio, Y. Tanaka, A. Kumanogoh, M. Traylor, C. M. Lewis, S. Eyre, H. Xu, R. Saxena, T. Arayssi, Y. Kochi, K. Ikari, M. Harigai, P. K. Gregersen, K. Yamamoto, S. Louis Bridges, Jr., L. Padyukov, J. Martin, L. Klareskog, Y. Okada, S. Raychaudhuri, Multi-ancestry genome-wide association analyses identify novel genetic mechanisms in rheumatoid arthritis. *Nat Genet* **54**, 1640-1651 (2022).

74. J. A. Morris, J. P. Kemp, S. E. Youlten, L. Laurent, J. G. Logan, R. C. Chai, N. A. Vulpescu, V. Forgetta, A. Kleinman, S. T. Mohanty, C. M. Sergio, J. Quinn, L. Nguyen-Yamamoto, A. L. Luco, J. Vijay, M. M. Simon, A. Pramatarova, C. Medina-Gomez, K. Trajanoska, E. J. Ghirardello, N. C. Butterfield, K. F. Curry, V. D. Leitch, P. C. Sparkes, A. T. Adoum, N. S. Mannan, D. S. K. Komla-Ebri, A. S. Pollard, H. F. Dewhurst, T. A. D. Hassall, M. G. Beltejar, T. andMe Research, D. J. Adams, S. M. Vaillancourt, S. Kaptoge, P. Baldock, C. Cooper, J. Reeve, E. E. Ntzani, E. Evangelou, C. Ohlsson, D. Karasik, F. Rivadeneira, D. P. Kiel, J. H. Tobias, C. L. Gregson, N. C. Harvey, E. Grundberg, D. Goltzman, D. J. Adams, C. J. Lelliott, D. A. Hinds, C. L. Ackert-Bicknell, Y. H. Hsu, M. T. Maurano, P. I. Croucher, G. R. Williams, J. H. D. Bassett, D. M. Evans, J. B. Richards, An atlas of genetic influences on osteoporosis in humans and mice. *Nat Genet* **51**, 258-266 (2019).

75. V. Trubetskoy, A. F. Pardinas, T. Qi, G. Panagiotaropoulou, S. Awasthi, T. B. Bigdeli, J. Bryois, C. Y. Chen, C. A. Dennison, L. S. Hall, M. Lam, K. Watanabe, O. Frei, T. Ge, J. C. Harwood, F. Koopmans, S. Magnusson, A. L. Richards, J. Sidorenko, Y. Wu, J. Zeng, J. Grove, M. Kim, Z. Li, G. Voloudakis, W. Zhang, M. Adams, I. Agartz, E. G. Atkinson, E. Agerbo, M. Al Eissa, M. Albus, M. Alexander, B. Z. Alizadeh, K. Alptekin, T. D. Als, F. Amin, V. Arolt, M. Arrojo, L. Athanasiu, M. H. Azevedo, S. A. Bacanu, N. J. Bass, M. Begemann, R. A. Belliveau, J. Bene, B. Benyamin, S. E. Bergen, G. Blasi, J. Bobes, S. Bonassi, A. Braun, R. A. Bressan, E. J. Bromet, R. Bruggeman, P. F. Buckley, R. L. Buckner, J. Bybjerg-Grauholm, W. Cahn, M. J. Cairns, M. E. Calkins, V. J. Carr, D. Castle, S. V. Catts, K. D. Chambert, R. C. K. Chan, B. Chaumette, W. Cheng, E. F. C. Cheung, S. A. Chong, D. Cohen, A. Consoli, Q. Cordeiro, J. Costas, C. Curtis, M. Davidson, K. L. Davis, L. de Haan, F. Degenhardt, L. E. DeLisi, D. Demontis, F. Dickerson, D. Dikeos, T. Dinan, S. Djurovic, J. Duan, G. Ducci, F. Dudbridge, J. G. Eriksson, L. Fananas, S. V. Faraone, A. Fiorentino, A. Forstner, J. Frank, N. B. Freimer, M. Fromer, A. Frustaci, A. Gadelha, G. Genovese, E. S. Gershon, M. Giannitelli, I. Giegling, P. Giusti-Rodriguez, S. Godard, J. I. Goldstein, J. Gonzalez Penas, A. Gonzalez-Pinto, S. Gopal, J. Gratten, M. F. Green, T. A. Greenwood, O. Guillin, S. Guloksuz, R. E. Gur, R. C. Gur, B. Gutierrez, E. Hahn, H. Hakonarson, V. Haroutunian, A. M. Hartmann, C. Harvey, C. Hayward, F. A. Henskens, S. Herms, P. Hoffmann, D. P. Howrigan, M. Ikeda, C. Iyegbe, I. Joa, A. Julia, A. K. Kahler, T. Kam-Thong, Y. Kamatani, S. Karachanak-Yankova, O. Kebir, M. C. Keller, B. J. Kelly, A. Khrunin, S. W. Kim, J. Klovins, N. Kondratiev, B. Konte, J. Kraft, M. Kubo, V. Kucinskas, Z. A. Kucinskiene, A. Kusumawardhani, H. Kuzelova-Ptackova, S. Landi, L. C. Lazzeroni, P. H. Lee, S. E. Legge, D. S. Lehrer, R. Lencer, B. Lerer, M. Li, J. Lieberman, G. A. Light, S. Limborska, C. M. Liu, J. Lonnqvist, C. M. Loughland, J. Lubinski, J. J. Luykx, A. Lynham, M. Macek, Jr., A. Mackinnon, P. K. E. Magnusson, B. S. Maher, W. Maier, D. Malaspina, J. Mallet, S. R. Marder, S. Marsal, A. R. Martin, L. Martorell, M. Mattheisen, R. W. McCarley, C. McDonald, J. J. McGrath, H. Medeiros, S. Meier, B. Melegh, I. Melle, R. I. Mesholam-Gately, A. Metspalu, P. T. Michie, L. Milani, V. Milanova, M. Mitjans, E. Molden, E. Molina, M. D. Molto, V. Mondelli, C. Moreno, C. P. Morley, G. Muntane, K. C. Murphy, I. Myin-Germeys, I. Nenadic, G. Nestadt, L. Nikitina-Zake, C. Noto, K. H. Nuechterlein, N. L. O'Brien, F. A. O'Neill, S. Y. Oh, A. Olincy, V. K. Ota, C. Pantelis, G.

N. Papadimitriou, M. Parellada, T. Paunio, R. Pellegrino, S. Periyasamy, D. O. Perkins, B. Pfuhlmann, O. Pietilainen, J. Pimm, D. Porteous, J. Powell, D. Quattrone, D. Quested, A. D. Radant, A. Rampino, M. H. Rapaport, A. Rautanen, A. Reichenberg, C. Roe, J. L. Roffman, J. Roth, M. Rothermundt, B. P. F. Rutten, S. Saker-Delye, V. Salomaa, J. Sanjuan, M. L. Santoro, A. Savitz, U. Schall, R. J. Scott, L. J. Seidman, S. I. Sharp, J. Shi, L. J. Siever, E. Sigurdsson, K. Sim, N. Skarabis, P. Slominsky, H. C. So, J. L. Sobell, E. Soderman, H. J. Stain, N. E. Steen, A. A. Steixner-Kumar, E. Stogmann, W. S. Stone, R. E. Straub, F. Streit, E. Strengman, T. S. Stroup, M. Subramaniam, C. A. Sugar, J. Suvisaari, D. M. Svrakic, N. R. Swerdlow, J. P. Szatkiewicz, T. M. T. Ta, A. Takahashi, C. Terao, F. Thibaut, D. Toncheva, P. A. Tooney, S. Torretta, S. Tosato, G. B. Tura, B. I. Turetsky, A. Ucok, A. Vaaler, T. van Amelsvoort, R. van Winkel, J. Veijola, J. Waddington, H. Walter, A. Waterreus, B. T. Webb, M. Weiser, N. M. Williams, S. H. Witt, B. K. Wormley, J. Q. Wu, Z. Xu, R. Yolken, C. C. Zai, W. Zhou, F. Zhu, F. Zimprich, E. C. Atbasoglu, M. Ayub, C. Benner, A. Bertolino, D. W. Black, N. J. Bray, G. Breen, N. G. Buccola, W. F. Byerley, W. J. Chen, C. R. Cloninger, B. Crespo-Facorro, G. Donohoe, R. Freedman, C. Galletly, M. J. Gandal, M. Gennarelli, D. M. Hougaard, H. G. Hwu, A. V. Jablensky, S. A. McCarroll, J. L. Moran, O. Mors, P. B. Mortensen, B. Muller-Myhsok, A. L. Neil, M. Nordentoft, M. T. Pato, T. L. Petryshen, M. Pirinen, A. E. Pulver, T. G. Schulze, J. M. Silverman, J. W. Smoller, E. A. Stahl, D. W. Tsuang, E. Vilella, S. H. Wang, S. Xu, C. Indonesia Schizophrenia, PsychEncode, C. Psychosis Endophenotypes International, G. O. C. Syn, R. Adolfsson, C. Arango, B. T. Baune, S. I. Belangero, A. D. Borglum, D. Braff, E. Bramon, J. D. Buxbaum, D. Campion, J. A. Cervilla, S. Cichon, D. A. Collier, A. Corvin, D. Curtis, M. D. Forti, E. Domenici, H. Ehrenreich, V. Escott-Price, T. Esko, A. H. Fanous, A. Gareeva, M. Gawlik, P. V. Gejman, M. Gill, S. J. Glatt, V. Golimbet, K. S. Hong, C. M. Hultman, S. E. Hyman, N. Iwata, E. G. Jonsson, R. S. Kahn, J. L. Kennedy, E. Khusnutdinova, G. Kirov, J. A. Knowles, M. O. Krebs, C. Laurent-Levinson, J. Lee, T. Lencz, D. F. Levinson, Q. S. Li, J. Liu, A. K. Malhotra, D. Malhotra, A. McIntosh, A. McQuillin, P. R. Menezes, V. A. Morgan, D. W. Morris, B. J. Mowry, R. M. Murray, V. Nimgaonkar, M. M. Nothen, R. A. Ophoff, S. A. Paciga, A. Palotie, C. N. Pato, S. Qin, M. Rietschel, B. P. Riley, M. Rivera, D. Rujescu, M. C. Saka, A. R. Sanders, S. G. Schwab, A. Serretti, P. C. Sham, Y. Shi, D. St Clair, H. Stefansson, K. Stefansson, M. T. Tsuang, J. van Os, M. P. Vawter, D. R. Weinberger, T. Werge, D. B. Wildenauer, X. Yu, W. Yue, P. A. Holmans, A. J. Pocklington, P. Roussos, E. Vassos, M. Verhage, P. M. Visscher, J. Yang, D. Posthuma, O. A. Andreassen, K. S. Kendler, M. J. Owen, N. R. Wray, M. J. Daly, H. Huang, B. M. Neale, P. F. Sullivan, S. Ripke, J. T. R. Walters, M. C. O'Donovan, C. Schizophrenia Working Group of the Psychiatric Genomics, Mapping genomic loci implicates genes and synaptic biology in schizophrenia. *Nature* **604**, 502-508 (2022).

76. A. R. Bentley, Y. J. Sung, M. R. Brown, T. W. Winkler, A. T. Kraja, I. Ntalla, K. Schwander, D. I. Chasman, E. Lim, X. Deng, X. Guo, J. Liu, Y. Lu, C. Y. Cheng, X. Sim, D. Vojinovic, J. E. Huffman, S. K. Musani, C. Li, M. F. Feitosa, M. A. Richard, R. Noordam, J. Baker, G. Chen, H. Aschard, T. M. Bartz, J. Ding, R. Dorajoo, A. K. Manning, T. Rankinen, A. V. Smith, S. M. Tajuddin, W. Zhao, M. Graff, M. Alver, M. Boissel, J. F. Chai, X. Chen, J. Divers, E. Evangelou, C. Gao, A. Goel, Y. Hagemeijer, S. E. Harris, F. P. Hartwig, M. He, A. Horimoto, F. C. Hsu, Y. J. Hung, A. U. Jackson, A. Kasturiratne, P. Komulainen, B. Kuhnel, K. Leander, K. H. Lin, J. Luan, L. P. Lyytikainen, N. Matoba, I. M. Nolte, M. Pietzner, B. Prins, M. Riaz, A. Robino, M. A. Said, N. Schupf, R. A. Scott, T. Sofer, A. Stancakova, F. Takeuchi, B. O. Tayo, P. J. van der Most, T. V. Varga, T. D. Wang, Y. Wang, E. B. Ware, W. Wen, Y. B. Xiang, L. R. Yanek, W. Zhang, J. H. Zhao, A. Adeyemo, S. Afaq, N. Amin, M. Amini, D. E. Arking,

Z. Arzumanyan, T. Aung, C. Ballantyne, R. G. Barr, L. F. Bielak, E. Boerwinkle, E. P. Bottinger, U. Broeckel, M. Brown, B. E. Cade, A. Campbell, M. Canouil, S. Charumathi, Y. I. Chen, K. Christensen, C. O.-K. Consortium, M. P. Concas, J. M. Connell, L. de las Fuentes, H. J. de Silva, P. S. de Vries, A. Doumatey, Q. Duan, C. B. Eaton, R. N. Eppinga, J. D. Faul, J. S. Floyd, N. G. Forouhi, T. Forrester, Y. Friedlander, I. Gandin, H. Gao, M. Ghanbari, S. A. Gharib, B. Gigante, F. Giulianini, H. J. Grabe, C. C. Gu, T. B. Harris, S. Heikkinen, C. K. Heng, M. Hirata, J. E. Hixson, M. A. Ikram, E. P.-I. Consortium, Y. Jia, R. Joehanes, C. Johnson, J. B. Jonas, A. E. Justice, T. Katsuya, C. C. Khor, T. O. Kilpelainen, W. P. Koh, I. Kolcic, C. Kooperberg, J. E. Krieger, S. B. Kritchevsky, M. Kubo, J. Kuusisto, T. A. Lakka, C. D. Langefeld, C. Langenberg, L. J. Launer, B. Lehne, C. E. Lewis, Y. Li, J. Liang, S. Lin, C. T. Liu, J. Liu, K. Liu, M. Loh, K. K. Lohman, T. Louie, A. Luzzi, R. Magi, A. Mahajan, A. W. Manichaikul, C. A. McKenzie, T. Meitinger, A. Metspalu, Y. Milaneschi, L. Milani, K. L. Mohlke, Y. Momozawa, A. P. Morris, A. D. Murray, M. A. Nalls, M. Nauck, C. P. Nelson, K. E. North, J. R. O'Connell, N. D. Palmer, G. J. Papanicolau, N. L. Pedersen, A. Peters, P. A. Peyser, O. Polasek, N. Poulter, O. T. Raitakari, A. P. Reiner, F. Renstrom, T. K. Rice, S. S. Rich, J. G. Robinson, L. M. Rose, F. R. Rosendaal, I. Rudan, C. O. Schmidt, P. J. Schreiner, W. R. Scott, P. Sever, Y. Shi, S. Sidney, M. Sims, J. A. Smith, H. Snieder, J. M. Starr, K. Strauch, H. M. Stringham, N. Y. Q. Tan, H. Tang, K. D. Taylor, Y. Y. Teo, Y. C. Tham, H. Tiemeier, S. T. Turner, A. G. Uitterlinden, G. Understanding Society Scientific, D. van Heemst, M. Waldenberger, H. Wang, L. Wang, L. Wang, W. B. Wei, C. A. Williams, G. Wilson, Sr., M. K. Wojczynski, J. Yao, K. Young, C. Yu, J. M. Yuan, J. Zhou, A. B. Zonderman, D. M. Becker, M. Boehnke, D. W. Bowden, J. C. Chambers, R. S. Cooper, U. de Faire, I. J. Deary, P. Elliott, T. Esko, M. Farrall, P. W. Franks, B. I. Freedman, P. Froguel, P. Gasparini, C. Gieger, B. L. Horta, J. J. Juang, Y. Kamatani, C. M. Kammerer, N. Kato, J. S. Kooner, M. Laakso, C. C. Laurie, I. T. Lee, T. Lehtimaki, C. Lifelines, P. K. E. Magnusson, A. J. Oldehinkel, B. Penninx, A. C. Pereira, R. Rauramaa, S. Redline, N. J. Samani, J. Scott, X. O. Shu, P. van der Harst, L. E. Wagenknecht, J. S. Wang, Y. X. Wang, N. J. Wareham, H. Watkins, D. R. Weir, A. R. Wickremasinghe, T. Wu, E. Zeggini, W. Zheng, C. Bouchard, M. K. Evans, V. Gudnason, S. L. R. Kardia, Y. Liu, B. M. Psaty, P. M. Ridker, R. M. van Dam, D. O. Mook-Kanamori, M. Fornage, M. A. Province, T. N. Kelly, E. R. Fox, C. Hayward, C. M. van Duijn, E. S. Tai, T. Y. Wong, R. J. F. Loos, N. Franceschini, J. I. Rotter, X. Zhu, L. J. Bierut, W. J. Gauderman, K. Rice, P. B. Munroe, A. C. Morrison, D. C. Rao, C. N. Rotimi, L. A. Cupples, Multi-ancestry genome-wide gene-smoking interaction study of 387,272 individuals identifies new loci associated with serum lipids. *Nat Genet* **51**, 636-648 (2019).

77. J. Waage, M. Standl, J. A. Curtin, L. E. Jessen, J. Thorsen, C. Tian, N. Schoettler, T. andMe Research, A. collaborators, C. Flores, A. Abdellaoui, T. S. Ahluwalia, A. C. Alves, A. F. S. Amaral, J. M. Anto, A. Arnold, A. Barreto-Luis, H. Baurecht, C. E. M. van Beijsterveldt, E. R. Bleecker, S. Bonas-Guarch, D. I. Boomsma, S. Brix, S. Bunyavanich, E. G. Burchard, Z. Chen, I. Curjuric, A. Custovic, H. T. den Dekker, S. C. Dharmage, J. Dmitrieva, L. Duijts, M. J. Ege, W. J. Gauderman, M. Georges, C. Gieger, F. Gilliland, R. Granell, H. Gui, T. Hansen, J. Heinrich, J. Henderson, N. Hernandez-Pacheco, P. Holt, M. Imboden, V. W. V. Jaddoe, M. R. Jarvelin, D. L. Jarvis, K. K. Jensen, I. Jonsdottir, M. Kabesch, J. Kaprio, A. Kumar, Y. A. Lee, A. M. Levin, X. Li, F. Lorenzo-Diaz, E. Melen, J. M. Mercader, D. A. Meyers, R. Myers, D. L. Nicolae, E. A. Nohr, T. Palviainen, L. Paternoster, C. E. Pennell, G. Pershagen, M. Pino-Yanes, N. M. Probst-Hensch, F. Ruschendorf, A. Simpson, K. Stefansson, J. Sunyer, G. Sveinbjornsson, E. Thiering, P. J. Thompson, M. Torrent, D. Torrents, J. Y. Tung, C. A. Wang, S. Weidinger, S. Weiss, G. Willemsen, L. K. Williams, C. Ober, D. A. Hinds, M. A. Ferreira, H. Bisgaard, D. P.

Strachan, K. Bonnelykke, Genome-wide association and HLA fine-mapping studies identify risk loci and genetic pathways underlying allergic rhinitis. *Nat Genet* **50**, 1072-1080 (2018).
78. D. T. Miller, K. Lee, N. S. Abul-Husn, L. M. Amendola, K. Brothers, W. K. Chung, M. H. Gollob, A. S. Gordon, S. M. Harrison, R. E. Hershberger, T. E. Klein, C. S. Richards, D. R. Stewart, C. L. Martin, A. S. F. W. G. E. a. documents@acmg.net, 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).
79. B. Zhao, T. Luo, T. Li, Y. Li, J. Zhang, Y. Shan, X. Wang, L. Yang, F. Zhou, Z. Zhu, I. Alzheimer's Disease Neuroimaging, N. Pediatric Imaging, Genetics, H. Zhu, Genome-wide association analysis of 19,629 individuals identifies variants influencing regional brain volumes and refines their genetic co-architecture with cognitive and mental health traits. *Nat Genet* **51**, 1637-1644 (2019).
80. C. J. Stevens, C. Murphy, R. Roberts, L. Lucas, F. Silva, D. Q. Fuller, Between China and South Asia: A Middle Asian corridor of crop dispersal and agricultural innovation in the Bronze Age. *Holocene* **26**, 1541-1555 (2016).
81. R. G. Allaby, C. Stevens, L. Lucas, O. Maeda, D. Q. Fuller, Geographic mosaics and changing rates of cereal domestication. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* **372**, (2017).
82. T. C. Zeng, L. A. Vyazov, A. Kim, P. Flegontov, K. Sirak, R. Maier, I. Lazaridis, A. Akbari, M. Frachetti, A. A. Tishkin, N. E. Ryabogina, S. A. Agapov, D. S. Agapov, A. N. Alekseev, G. G. Boeskorov, A. P. Derevianko, V. M. Dyakonov, D. N. Enshin, A. V. Fribus, Y. V. Frolov, S. P. Grushin, A. A. Khokhlov, K. Y. Kiryushin, Y. F. Kiryushin, E. P. Kitov, P. Kosintsev, I. V. Kovtun, N. P. Makarov, V. V. Morozov, E. N. Nikolaev, M. P. Rykun, T. M. Savenkova, M. V. Shchelchkova, V. Shirokov, S. N. Skochina, O. S. Sherstobitova, S. M. Slepchenko, K. N. Solodovnikov, E. N. Solovyova, A. D. Stepanov, A. A. Timoshchenko, A. S. Vdovin, A. V. Vybornov, E. V. Balanovska, S. Dryomov, G. Hellenthal, K. Kidd, J. Krause, E. Starikovskaya, R. Sukernik, T. Tatarinova, M. G. Thomas, M. Zhabagin, K. Callan, O. Cheronet, D. Fernandes, D. Keating, F. Candilio, L. Iliev, A. Kearns, K. T. Ozdogan, M. Mah, A. Micco, M. Michel, I. Olalde, F. Zalzala, S. Mallick, N. Rohland, R. Pinhasi, V. M. Narasimhan, D. Reich, Ancient DNA reveals the prehistory of the Uralic and Yeniseian peoples. *Nature* **644**, 122-132 (2025).
83. H. Matsumae, P. Ranacher, P. E. Savage, D. E. Blasi, T. E. Currie, K. Koganebuchi, N. Nishida, T. Sato, H. Tanabe, A. Tajima, S. Brown, M. Stoneking, K. K. Shimizu, H. Oota, B. Bickel, Exploring correlations in genetic and cultural variation across language families in northeast Asia. *Sci Adv* **7**, eabd9223 (2021).
84. M. Robbeets, R. Bouckaert, M. Conte, A. Savelyev, T. Li, D. I. An, K. I. Shinoda, Y. Cui, T. Kawashima, G. Kim, J. Uchiyama, J. Dolinska, S. Oskolskaya, K. Y. Yamano, N. Seguchi, H. Tomita, H. Takamiya, H. Kanzawa-Kiriyama, H. Oota, H. Ishida, R. Kimura, T. Sato, J. H. Kim, B. Deng, R. Bjorn, S. Rhee, K. D. Ahn, I. Gruntov, O. Mazo, J. R. Bentley, R. Fernandes, P. Roberts, I. R. Bausch, L. Gilaizeau, M. Yoneda, M. Kugai, R. A. Bianco, F. Zhang, M. Himmel, M. J. Hudson, C. Ning, Triangulation supports agricultural spread of the Transeurasian languages. *Nature* **599**, 616-621 (2021).
85. P. B. Golden, C. Hriban, *Studies on the Peoples and Cultures of the Eurasian Steppes* (Ed. Acad. Române, 2011).
86. G. He, A. Adnan, A. Rakha, H. Y. Yeh, M. Wang, X. Zou, J. Guo, M. Rehman, A. Fawad, P. Chen, C. C. Wang, A comprehensive exploration of the genetic legacy and forensic

features of Afghanistan and Pakistan Mongolian-descent Hazara. *Forensic Sci Int Genet* **42**, e1-e12 (2019).
87. X. Ma, W. Yang, Y. Gao, Y. Pan, Y. Lu, H. Chen, D. Lu, S. Xu, Genetic Origins and Sex-Biased Admixture of the Huis. *Mol Biol Evol* **38**, 3804-3819 (2021).
88. W. Barrie, Y. Yang, E. K. Irving-Pease, K. E. Attfield, G. Scorrano, L. T. Jensen, A. P. Armen, E. A. Dimopoulos, A. Stern, A. Refoyo-Martinez, A. Pearson, A. Ramsoe, C. Gaunitz, F. Demeter, M. L. S. Jorkov, S. B. Moller, B. Springborg, L. Klassen, I. M. Hyldgard, N. Wickmann, L. Vinner, T. S. Korneliussen, M. E. Allentoft, M. Sikora, K. Kristiansen, S. Rodriguez, R. Nielsen, A. K. N. Iversen, D. J. Lawson, L. Fugger, E. Willerslev, Elevated genetic risk for multiple sclerosis emerged in steppe pastoralist populations. *Nature* **625**, 321-328 (2024).
89. P. Akbari, D. Vuckovic, L. Stefanucci, T. Jiang, K. Kundu, R. Kreuzhuber, E. L. Bao, J. H. Collins, K. Downes, L. Grassi, J. A. Guerrero, S. Kaptoge, J. C. Knight, S. Meacham, J. Sambrook, D. Seyres, O. Stegle, J. M. Verboon, K. Walter, N. A. Watkins, J. Danesh, D. J. Roberts, E. Di Angelantonio, V. G. Sankaran, M. Frontini, S. Burgess, T. Kuijpers, J. E. Peters, A. S. Butterworth, W. H. Ouwehand, N. Soranzo, W. J. Astle, A genome-wide association study of blood cell morphology identifies cellular proteins implicated in disease aetiology. *Nat Commun* **14**, 5023 (2023).
90. W. M. A. J. Jama, World Medical Association Declaration of Helsinki: ethical principles for medical research involving human subjects. **310**, 2191-2194 (2013).
91. K. I. Kendig, S. Baheti, M. A. Bockol, T. M. Drucker, S. N. Hart, J. R. Heldenbrand, M. Hernaez, M. E. Hudson, M. T. Kalmbach, E. W. Klee, N. R. Mattson, C. A. Ross, M. Taschuk, E. D. Wieben, M. Wiepert, D. E. Wildman, L. S. Mainzer, Sentieon DNASeq Variant Calling Workflow Demonstrates Strong Computational Performance and Accuracy. *Front Genet* **10**, 736 (2019).
92. H. Li, R. Durbin, Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754-1760 (2009).
93. H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, S. Genome Project Data Processing, The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079 (2009).
94. A. McKenna, M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, K. Garimella, D. Altshuler, S. Gabriel, M. Daly, M. A. DePristo, The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* **20**, 1297-1303 (2010).
95. C. C. Chang, C. C. Chow, L. C. Tellier, S. Vattikuti, S. M. Purcell, J. J. Lee, Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**, 7 (2015).
96. A. Manichaikul, J. C. Mychaleckyj, S. S. Rich, K. Daly, M. Sale, W. M. Chen, Robust relationship inference in genome-wide association studies. *Bioinformatics* **26**, 2867-2873 (2010).
97. A. Bergstrom, S. A. McCarthy, R. Hui, M. A. Almarri, Q. Ayub, P. Danecek, Y. Chen, S. Felkel, P. Hallast, J. Kamm, H. Blanche, J. F. Deleuze, H. Cann, S. Mallick, D. Reich, M. S. Sandhu, P. Skoglund, A. Scally, Y. Xue, R. Durbin, C. Tyler-Smith, Insights into human genetic variation and population history from 929 diverse genomes. *Science* **367**, (2020).
98. N. Patterson, A. L. Price, D. Reich, Population structure and eigenanalysis. *PLoS Genet* **2**, e190 (2006).
99. D. H. Alexander, J. Novembre, K. Lange, Fast model-based estimation of ancestry in unrelated individuals. *Genome Res* **19**, 1655-1664 (2009).

100. J. K. Pickrell, J. K. Pritchard, Inference of population splits and mixtures from genome-wide allele frequency data. *PLoS Genet* **8**, e1002967 (2012).
101. P. Dixon, VEGAN, a package of R functions for community ecology. *J Veg Sci* **14**, 927-930 (2003).
102. E. G. Atkinson, S. Dalvie, Y. Pichkar, A. Kalungi, L. Majara, A. Stevenson, T. Abebe, D. Akena, M. Alemayehu, F. K. Ashaba, L. Atwoli, M. Baker, L. B. Chibnik, N. Creanza, M. J. Daly, A. Fekadu, B. Gelaye, S. Gichuru, W. E. Injera, R. James, S. M. Kariuki, G. Kigen, N. Koen, K. C. Koenen, Z. Koenig, E. Kwobah, J. Kyebuzibwa, H. Musinguzi, R. M. Mwema, B. M. Neale, C. P. Newman, C. Newton, L. Ongeri, S. Ramachandran, R. Ramesar, W. Shiferaw, D. J. Stein, R. E. Stroud, S. Teferra, M. T. Yohannes, Z. Zingela, A. R. Martin, G. A. P. P. S. T. Neuro, Genetic structure correlates with ethnolinguistic diversity in eastern and southern Africa. *Am J Hum Genet* **109**, 1667-1679 (2022).
103. N. Patterson, P. Moorjani, Y. Luo, S. Mallick, N. Rohland, Y. Zhan, T. Genschoreck, T. Webster, D. Reich, Ancient admixture in human history. *Genetics* **192**, 1065-1093 (2012).
104. P. R. Loh, M. Lipson, N. Patterson, P. Moorjani, J. K. Pickrell, D. Reich, B. Berger, Inferring admixture histories of human populations using linkage disequilibrium. *Genetics* **193**, 1233-1254 (2013).
105. H. Chen, Y. Lu, D. Lu, S. Xu, Y-LineageTracker: a high-throughput analysis framework for Y-chromosomal next-generation sequencing data. *BMC Bioinformatics* **22**, 114 (2021).
106. A. Jagadeesan, S. S. Ebenesersdottir, V. B. Guethmundsdottir, E. L. Thordardottir, K. H. S. Moore, A. Helgason, HaploGrouper: a generalized approach to haplogroup classification. *Bioinformatics* **37**, 570-572 (2021).
107. H. Weissensteiner, D. Pacher, A. Kloss-Brandstatter, L. Forer, G. Specht, H. J. Bandelt, F. Kronenberg, A. Salas, S. Schonherr, HaploGrep 2: mitochondrial haplogroup classification in the era of high-throughput sequencing. *Nucleic Acids Res* **44**, W58-63 (2016).
108. I. Mathieson, S. Alpaslan-Roodenberg, C. Posth, A. Szecsenyi-Nagy, N. Rohland, S. Mallick, I. Olalde, N. Broomandkhoshbacht, F. Candilio, O. Cheronet, D. Fernandes, M. Ferry, B. Gamarra, G. G. Fortes, W. Haak, E. Harney, E. Jones, D. Keating, B. Krause-Kyora, I. Kucukkalipci, M. Michel, A. Mittnik, K. Nagele, M. Novak, J. Oppenheimer, N. Patterson, S. Pfrengle, K. Sirak, K. Stewardson, S. Vai, S. Alexandrov, K. W. Alt, R. Andreescu, D. Antonovic, A. Ash, N. Atanassova, K. Bacvarov, M. B. Gusztav, H. Bocherens, M. Bolus, A. Boroneant, Y. Boyadzhiev, A. Budnik, J. Burmaz, S. Chohadzhiev, N. J. Conard, R. Cottiaux, M. Cuka, C. Cupillard, D. G. Drucker, N. Elenski, M. Francken, B. Galabova, G. Ganetsovski, B. Gely, T. Hajdu, V. Handzhyiska, K. Harvati, T. Higham, S. Iliev, I. Jankovic, I. Karavanic, D. J. Kennett, D. Komso, A. Kozak, D. Labuda, M. Lari, C. Lazar, M. Leppek, K. Leshtakov, D. L. Vetro, D. Los, I. Lozanov, M. Malina, F. Martini, K. McSweeney, H. Meller, M. Mendusic, P. Mirea, V. Moiseyev, V. Petrova, T. D. Price, A. Simalcsik, L. Sineo, M. Slaus, V. Slavchev, P. Stanev, A. Starovic, T. Szeniczey, S. Talamo, M. Teschler-Nicola, C. Thevenet, I. Valchev, F. Valentin, S. Vasilyev, F. Veljanovska, S. Venelinova, E. Veselovskaya, B. Viola, C. Virag, J. Zaninovic, S. Zauner, P. W. Stockhammer, G. Catalano, R. Krauss, D. Caramelli, G. Zarina, B. Gaydarska, M. Lillie, A. G. Nikitin, I. Potekhina, A. Papathanasiou, D. Boric, C. Bonsall, J. Krause, R. Pinhasi, D. Reich, The genomic history of southeastern Europe. *Nature* **555**, 197-203 (2018).
109. O. Delaneau, J. Marchini, J. F. Zagury, A linear complexity phasing method for thousands of genomes. *Nat Methods* **9**, 179-181 (2011).
110. B. L. Browning, S. R. Browning, Improving the accuracy and efficiency of identity-by-descent detection in population data. *Genetics* **194**, 459-471 (2013).

111. D. J. Lawson, G. Hellenthal, S. Myers, D. Falush, Inference of population structure using dense haplotype data. *PLoS Genet* **8**, e1002453 (2012).
112. P. Wangkumhang, M. Greenfield, G. Hellenthal, An efficient method to identify, date, and describe admixture events using haplotype information. *Genome Res* **32**, 1553-1564 (2022).
113. J. Terhorst, J. A. Kamm, Y. S. Song, Robust and scalable inference of population history from hundreds of unphased whole genomes. *Nat Genet* **49**, 303-309 (2017).
114. S. Schiffels, K. Wang, MSMC and MSMC2: The Multiple Sequentially Markovian Coalescent. *Methods in molecular biology (Clifton, N.J.)* **2090**, 147-166 (2020).
115. L. Excoffier, I. Dupanloup, E. Huerta-Sanchez, V. C. Sousa, M. Foll, Robust demographic inference from genomic and SNP data. *PLoS Genet* **9**, e1003905 (2013).
116. W. McLaren, L. Gil, S. E. Hunt, H. S. Riat, G. R. Ritchie, A. Thormann, P. Flicek, F. Cunningham, The Ensembl Variant Effect Predictor. *Genome Biol* **17**, 122 (2016).
117. W. Aamer, A. Al-Maraghi, N. Syed, G. D. Gandhi, E. Aliyev, A. A. Al-Kurbi, O. Al-Saei, M. Kohailan, N. Krishnamoorthy, S. Palaniswamy, K. Al-Malki, S. Abbasi, N. Agrebi, F. Abbaszadeh, A. S. A. Akil, R. Badii, T. Ben-Omran, B. Lo, C. Qatar Genome Program Research, Y. Mokrab, K. A. Fakhro, Burden of Mendelian disorders in a large Middle Eastern biobank. *Genome Med* **16**, 46 (2024).
118. P. Cingolani, A. Platts, L. Wang le, M. Coon, T. Nguyen, L. Wang, S. J. Land, X. Lu, D. M. Ruden, A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly (Austin)* **6**, 80-92 (2012).
119. J. Wen, J. Liu, Q. Feng, Y. Lu, K. Yuan, X. Zhang, C. Zhang, Y. Gao, X. Wang, D. Mamatyusupu, S. Xu, Ancestral origins and post-admixture adaptive evolution of highland Tajiks. *Natl Sci Rev* **11**, nwae284 (2024).
120. Y. Zhou, B. Zhou, L. Pache, M. Chang, A. H. Khodabakhshi, O. Tanaseichuk, C. Benner, S. K. Chanda, Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. *Nat Commun* **10**, 1523 (2019).
121. S. R. Browning, B. L. Browning, Y. Zhou, S. Tucci, J. M. Akey, Analysis of Human Sequence Data Reveals Two Pulses of Archaic Denisovan Admixture. *Cell* **173**, 53-61 e59 (2018).
122. X. Liu, S. Koyama, K. Tomizuka, S. Takata, Y. Ishikawa, S. Ito, S. Kosugi, K. Suzuki, K. Hikino, M. Koido, Y. Koike, M. Horikoshi, T. Gakuhari, S. Ikegawa, K. Matsuda, Y. Momozawa, K. Ito, Y. Kamatani, C. Terao, Decoding triancestral origins, archaic introgression, and natural selection in the Japanese population by whole-genome sequencing. *Sci Adv* **10**, eadi8419 (2024).
123. A. R. Quinlan, I. M. Hall, BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841-842 (2010).
124. Z. A. Szpiech, R. D. Hernandez, selscan: an efficient multithreaded program to perform EHH-based scans for positive selection. *Mol Biol Evol* **31**, 2824-2827 (2014).

Acknowledgments

We thank Prof. Etienne Patin and Prof. Lluis Quintana-Murci from the Human Evolutionary Genetics Unit of the Institut Pasteur for sharing the high-coverage genomes of 317 individuals from the Pacific region. We also thank all the members of the Consortium of the Central Asian Genomic Diversity Project and Prof. Yerlan Ramankulov from Nazarbayev University for their support and consultation.

Funding: National Natural Science Foundation of China grant 82402203 (GLH)

Science Committee of the Ministry of Education and Science of the Republic of Kazakhstan grant BR18574101 (MZ)
National Natural Science Foundation of China grant 82202078 (MGW)
Faculty Development Competitive Research Grants Programs of Nazarbayev University grant SST2019012 (MZ)
Major Project of the National Social Science Foundation of China grant 23&ZD203 (LHW, GLH)
Open Project of the Key Laboratory of Forensic Genetics of the Ministry of Public Security grant 2022FGKFKT05 (GLH)
Open Project of the Key Laboratory of Forensic Genetics of the Ministry of Public Security grant 2024FGKFKT02 (MGW)
Center for Archaeological Science of Sichuan University grant 23SASA01 (GLH)
Center for Archaeological Science of Sichuan University grant 24SASB03 (GLH)
Sichuan Science and Technology Program grant 2024NSFSC1518 (GLH)

Author contributions:
Conceptualization: MGW, GLH, LHW, MZ
Methodology: MZ
Software: HRS, QXS, QXY, LTL, JZ
Validation: GLH, MGW, QXS, RKT, JZ, ZS
Formal analysis: MGW, GLH, HRS, QXS, QXY, LTL, JZ
Investigation: MGW, GLH, ZS, JC, FXB, YL, CL, HJY
Resources: GLH, MGW, LHW, MZ
Data curation: GLH, MGW, MZ
Writing - original draft: MGW, GLH
Writing - review & editing: MGW, GLH, RKT, ZS, CL, LHW, MZ
Visualization: MGW, GLH, HRS, QXS, QXY, LTL
Supervision: MGW, GLH, JC, FXB, YL, CL, HJY
Project administration: GLH, MGW, LHW, MZ
Funding acquisition: GLH, MGW, LHW, MZ

Competing interests: The authors declare that they have no competing interests.

Data and materials availability: The raw data of 166 individuals were submitted to the Genome Variation Map database under accession number GVM000900 (https://bigd.big.ac.cn/gvm/getProjectDetail?Project=GVM000900). All the data used here are included in the supplementary materials. Reference populations can be found in publicly available datasets, such as the Human Genetic Diversity Project dataset, Oceania genomic resource, and the Allen Ancient DNA Resource (HO and 1240K datasets) from the David Reich Lab (https://reich.hms.harvard.edu/datasets).

Figures and Tables

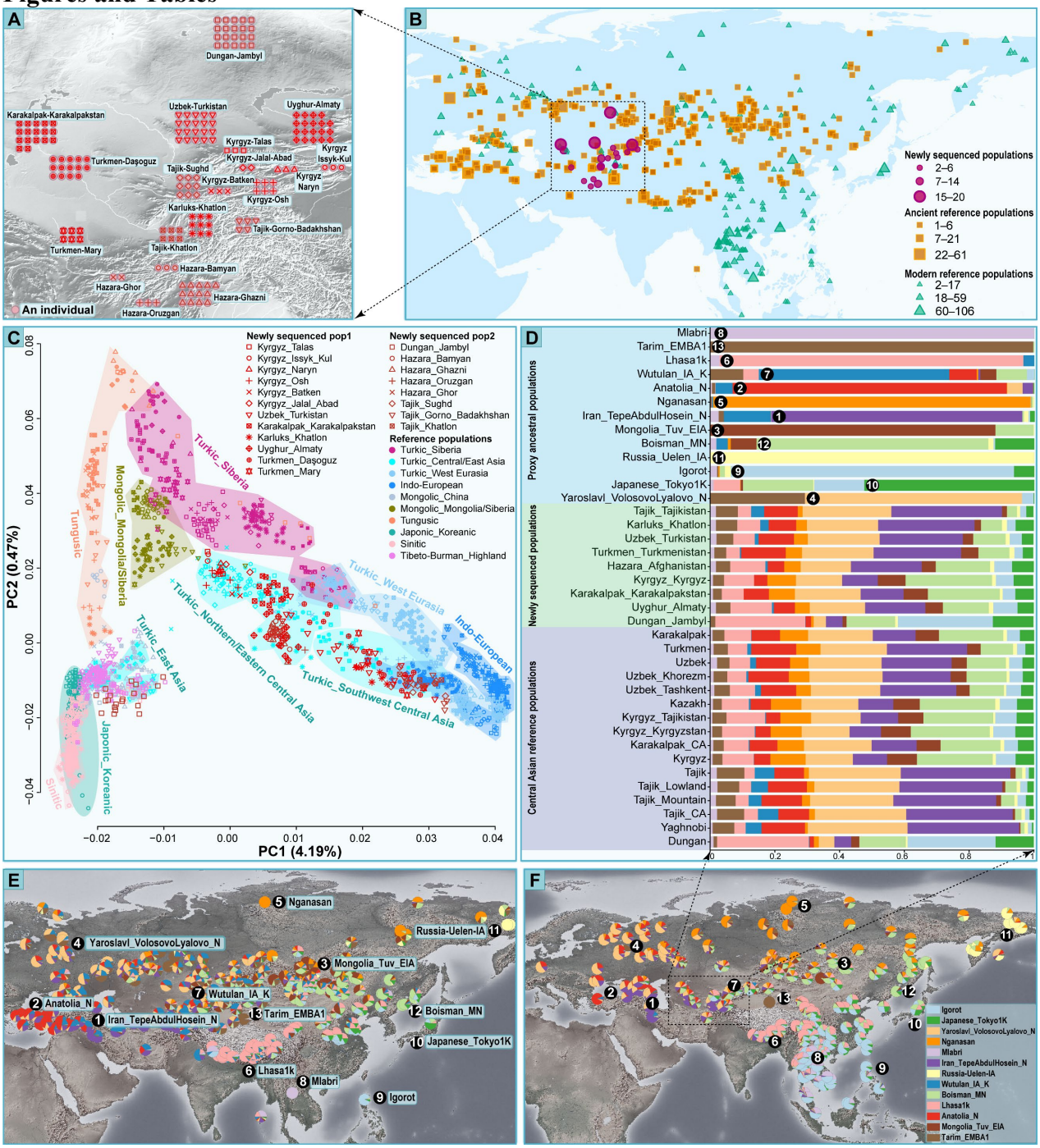


Fig. 1. Geographic distribution and population structure of newly sequenced Central Asian and Afghanistan Hazaras (CAAH) populations and reference populations based on the merged Human Origins (HO) dataset. (**A**), Sampling locations of 20 newly collected populations. Each circle represents a single individual. (**B**), Geographic distribution of newly sequenced populations alongside modern and ancient Eurasian reference populations from the extended HO dataset. The purple circles indicate newly sequenced populations, the orange squares represent ancient references, and the turquoise triangles denote modern references. (**C**), Patterns of genetic affinity between CAAH and 132 linguistically related Eurasian populations from the Altaic, Indo-European, and Sino-Tibetan language families. Genetic clusters or clines are indicated, with background and label colors corresponding to the groupings shown in fig. S2. Reference populations are presented by language families, with square colors preceding each group matching those in fig. S2. The specific populations comprising each group are detailed in

fig. S2. (**D**), Model-based ADMIXTURE analysis at K = 13, displaying the admixture profiles of Central Asian populations and proxy ancestral groups. Groups 1–13 correspond to proxy ancestral populations, with group numbers assigned arbitrarily. The colors denote distinct ancestral components. (**E**), Detailed admixture profiles of ancient Eurasian populations at K = 13. (**F**), Admixture profiles of modern Eurasian populations at K = 13.

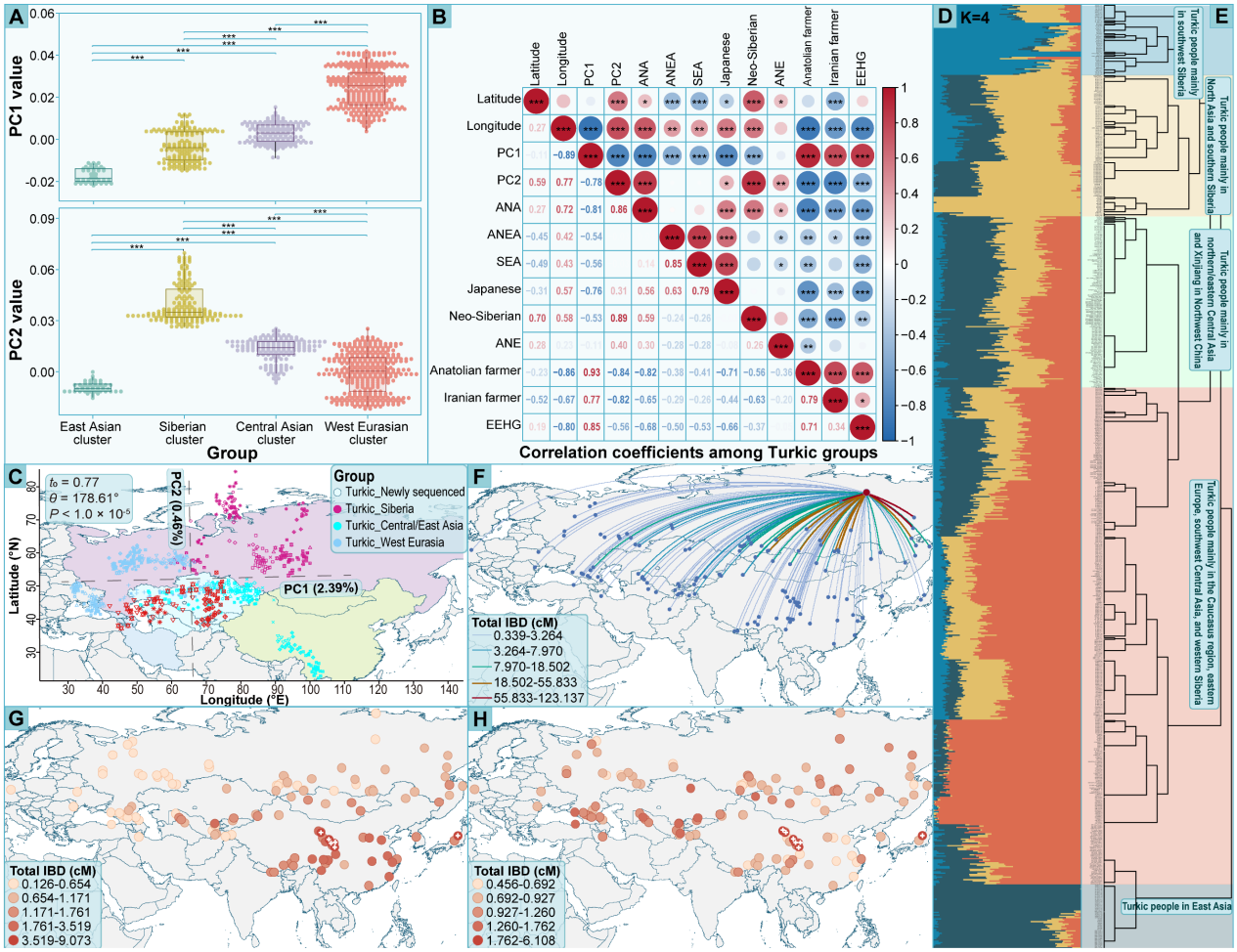


Fig. 2. Geographic and genetic correlations, fine-scale structure of Turkic groups, and identity-by-descent (IBD) sharing among Eurasian populations. (**A**), Pairwise Welch's t-tests of the first two PCs among four CAAH subgroups, as defined in Fig. 1C. Turkic-speaking populations from southwestern Central Asia and western Siberia were assigned to the West Eurasian cluster. "***" indicates $p < 0.001$. (**B**), Correlations between geographic coordinates (latitude, longitude) and genetic features (PC1, PC2, and ancestral components) among 54 Turkic groups. The lower left panel displays correlation coefficients (r), with negative (blue), positive (red), and neutral (0) values. The upper right panel presents p-values, with circle color (blue to red) and size reflecting the r-value. Significance levels: $0.01 \leq p < 0.05$ (*), $0.001 \leq p < 0.01$ (**), and $p < 0.001$ (***). ANA: ancient Northeast Asian ancestry; ANEA: ancient northern East Asian ancestry; SEA: southern East Asian ancestry; ANE: ancient North Eurasian ancestry; EEHG: Eastern European hunter-gatherers. (**C**), Procrustes-transformed PCA demonstrates a strong correlation between genetic differentiation and geography among 54 Turkic groups ($t_0 = 0.77$, $p < 1.0 \times 10^{-5}$, rotation angle $\theta = 178.61°$). Populations are grouped by geography with colors and shapes consistent with Fig. 1C and fig. S2. Shaded backgrounds denote the country origins of the samples. (**D**), ADMIXTURE results for Turkic-speaking populations at K = 4. (**E**), FineSTRUCTURE-based topology reveals fine-scale substructures among Turkic groups. (**F–H**), Total IBD sharing between the easternmost Turkic group (**F**), Jambyl Dungan (**G**), and

Afghanistan Hazara (**H**) with Eurasian reference populations in the low-density HO dataset. References include publicly available Altaic, Sinitic, Tibeto-Burman, and Indo-European groups. Circle positions follow the geographic distributions in fig. S12A; colors denote IBD length ranges shared with the integrated easternmost Turkic group (Dolgan and Yakut).

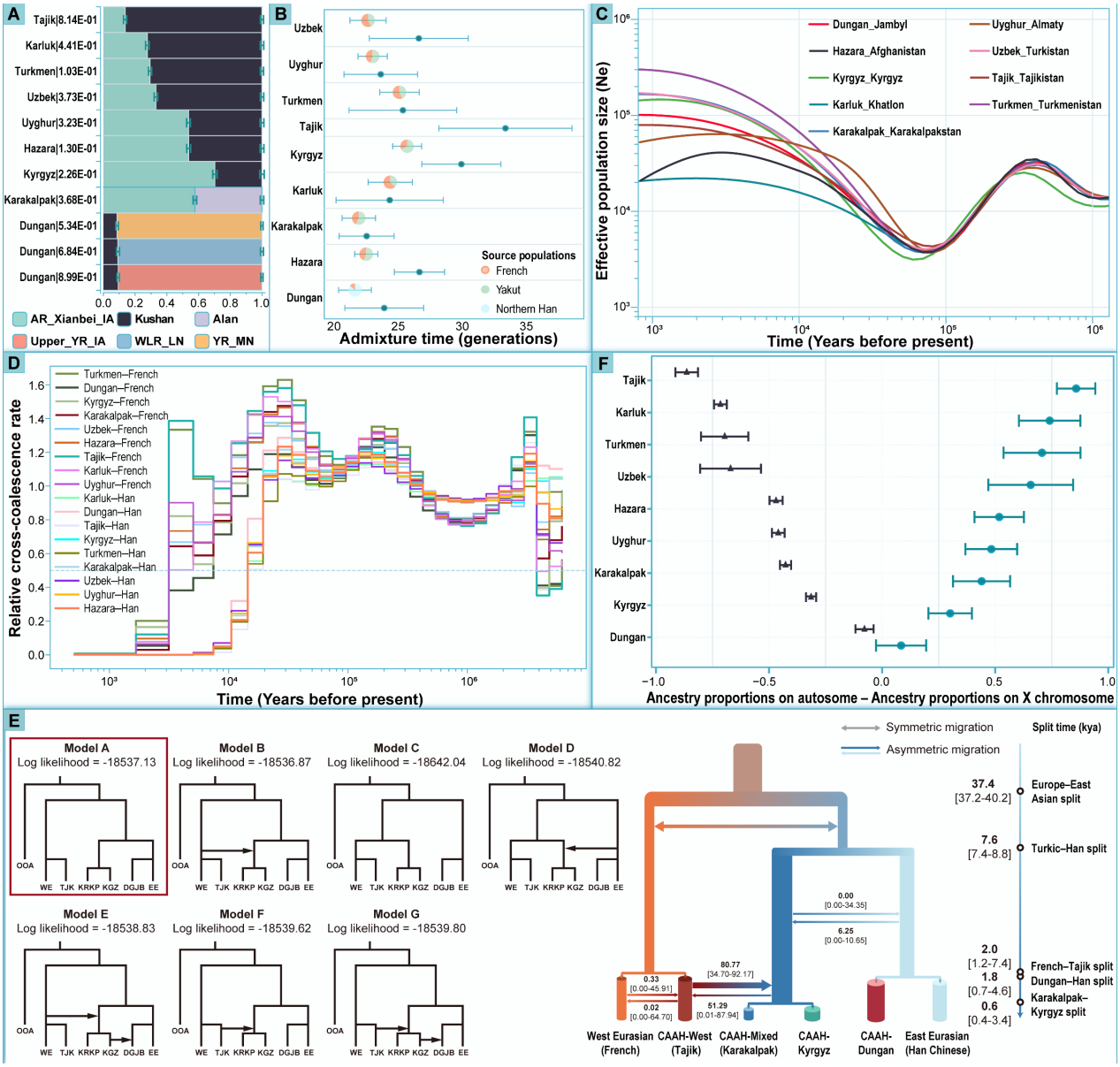


Fig. 3. Complex demographic modeling of CAAH populations. (**A**), Admixture proportions of each CAAH population based on qpAdm estimation. The numerical value following each population represents the *p*-value from the qpAdm analysis. A *p*-value > 0.05 indicates that the target population can be successfully modeled as a mixture of the specified source populations. (**B**), The admixture proportions were inferred from fastGLOBETROTTER analysis, and admixture times were inferred from fastGLOBETROTTER and MALDER analyses. The blue circles denote the MALDER-based admixture times. Other colored circles denote the fastGLOBETROTTER-based admixture times and ancestry compositions. The color of the circle represents the different surrogate ancestral sources. The number of admixtures is presented as the mean ± standard deviation. (**C**), Effective population sizes of each CAAH population. (**D**), The population divergence times between each CAAH population and East/West Eurasian source populations. EE: East Eurasian population, WE: West Eurasian population. (**E**), Sex-biased admixture patterns in CAAH inferred from autosomal and X-chromosomal variants. The

estimated proportions of western Eurasian ancestry are shown, with autosomal-derived values plotted as negative values. (**F**), Maximum-likelihood demographic model of Central Asian populations. Model A was selected as the optimal model despite a slightly lower likelihood than Model B, as the difference was not statistically significant and Model A required fewer parameters. Brackets indicate 95% confidence intervals obtained by non-parametric block bootstrap. Migration rates scaled by effective population size (2Nm) are shown above or below the corresponding arrows; arrows representing gene flow with 2Nm > 1 are bolded. OOA: Out of Africa; TJK: Tajik_Tajikistan; DGJB: Dungan_Jambyl; KRKP: Karakalpak_Karakalpakstan; KGZ: Kyrgyz_Kyrgyzstan.

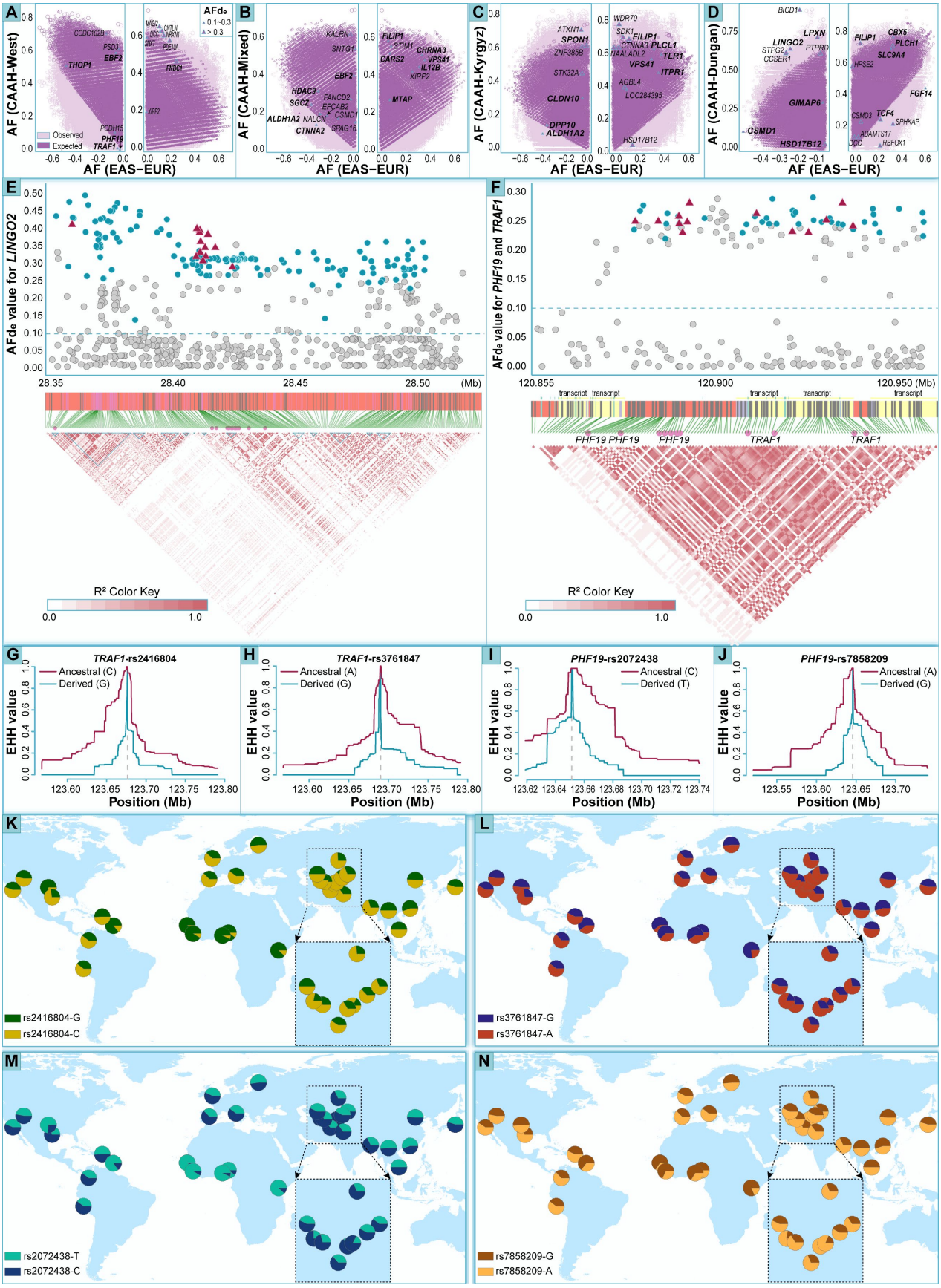


Fig. 4. The signatures of biological adaptation identified in genetically diverse CAAH groups and the frequency distribution patterns of specific variants. (**A–D**), Distribution of observed and expected allele frequencies (AFs) for variants within each CAAH in the context of

AF differences between East and West Eurasian populations. The top 10 genes per population, previously unreported in the GWAS Catalog for phenotypic effects, are shown in black, whereas well-studied genes are highlighted in red. (**E–F**), Linkage patterns and deviations between observed and expected allele frequencies (AFd_e) of variants in the *LINGO2* (**E**) and *TRAF1*/*PHF19* genes (**F**). The colored markers indicate natural selection signatures on these genes. The turquoise dots represent variants not previously associated with phenotypic effects in the GWAS Catalog, whereas the claret triangles denote variants that have been reported to exhibit phenotypic effects in the GWAS Catalog. (**G–J**), Extended haplotype homozygosity (EHH) curves for *TRAF1*-rs2416804 (**G**), *TRAF1*-rs3761847 (**H**), *PHF19*-rs2072438 (**I**), and *PHF19*-rs7858209 (**J**). (**K–N**), AF distributions of *TRAF1*-rs2416804 (**K**), *TRAF1*-rs3761847 (**L**), *PHF19*-rs2072438 (**M**), and *PHF19*-rs7858209 (**N**).

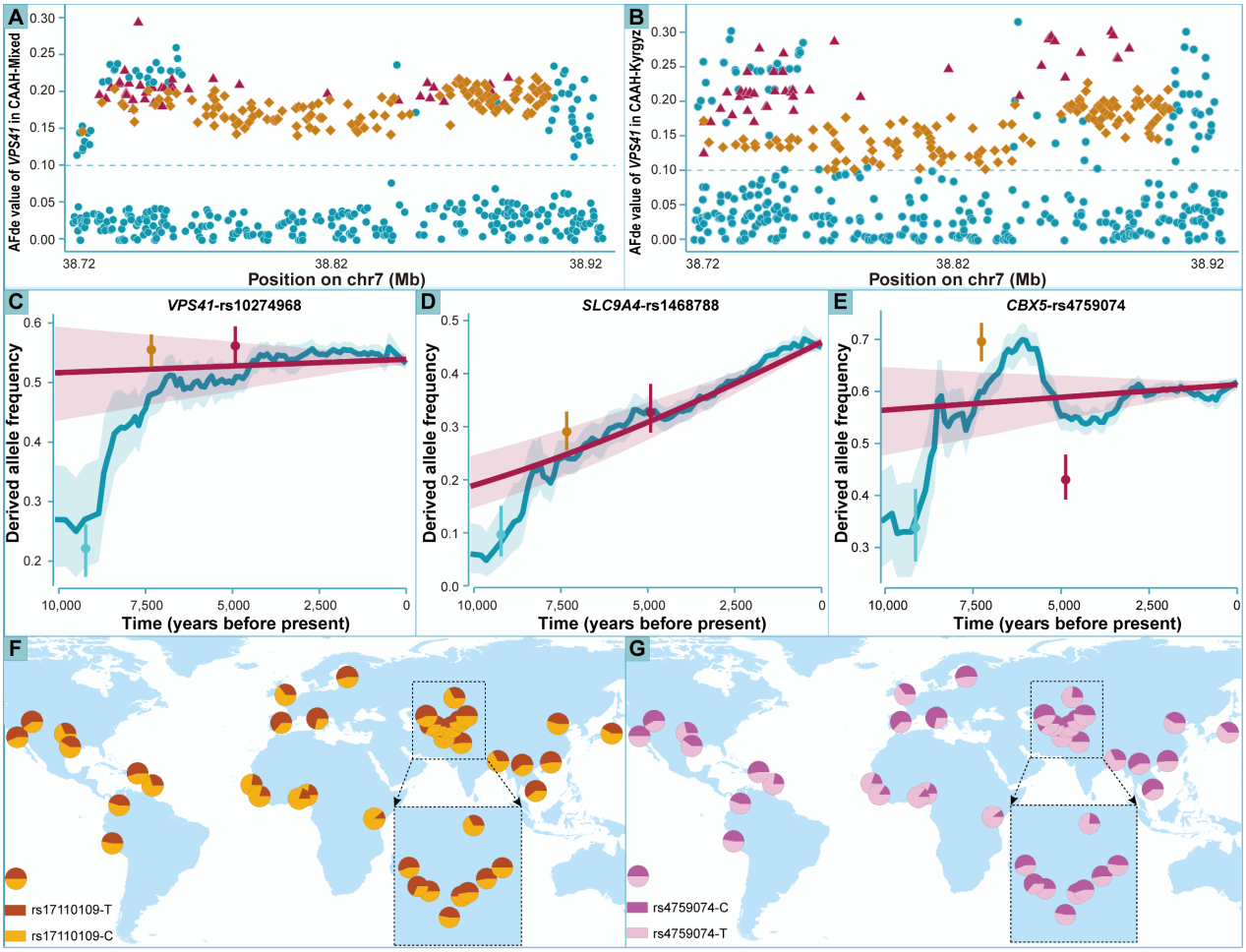


Fig. 5. AFd_e distribution of *VPS41* variants, derived allele frequency (DAF) trajectories of local adaptive signatures, and distribution of Neanderthal-derived archaic introgression segments (AISs) across Eurasian groups. (**A–B**), AFd_e distribution of *VPS41* variants in CAAH-Mixed (**A**) and CAAH-Kyrgyz (**B**). Loci marked in orange-yellow represent highly differentiated variants (HDVs) within *VPS41*, exhibiting a West Eurasian-biased frequency in CAAH-Mixed or CAAH-Kyrgyz. The claret-marked loci indicate HDVs with East Eurasian-biased frequencies in CAAH-Mixed or CAAH-Kyrgyz. Loci marked in turquoise denote variants within the *VPS41* gene that lack signals of natural selection. (**C–E**), DAF trajectories for *VPS41*-rs10274968 (**C**), *SLC9A4*-rs1468788 (**D**), and *CBX5*-rs4759074 (**E**). (**F–G**), AF distributions of *CBX5*-rs17110109 (**F**) and *CBX5*-rs4759074 (**G**) across global populations.

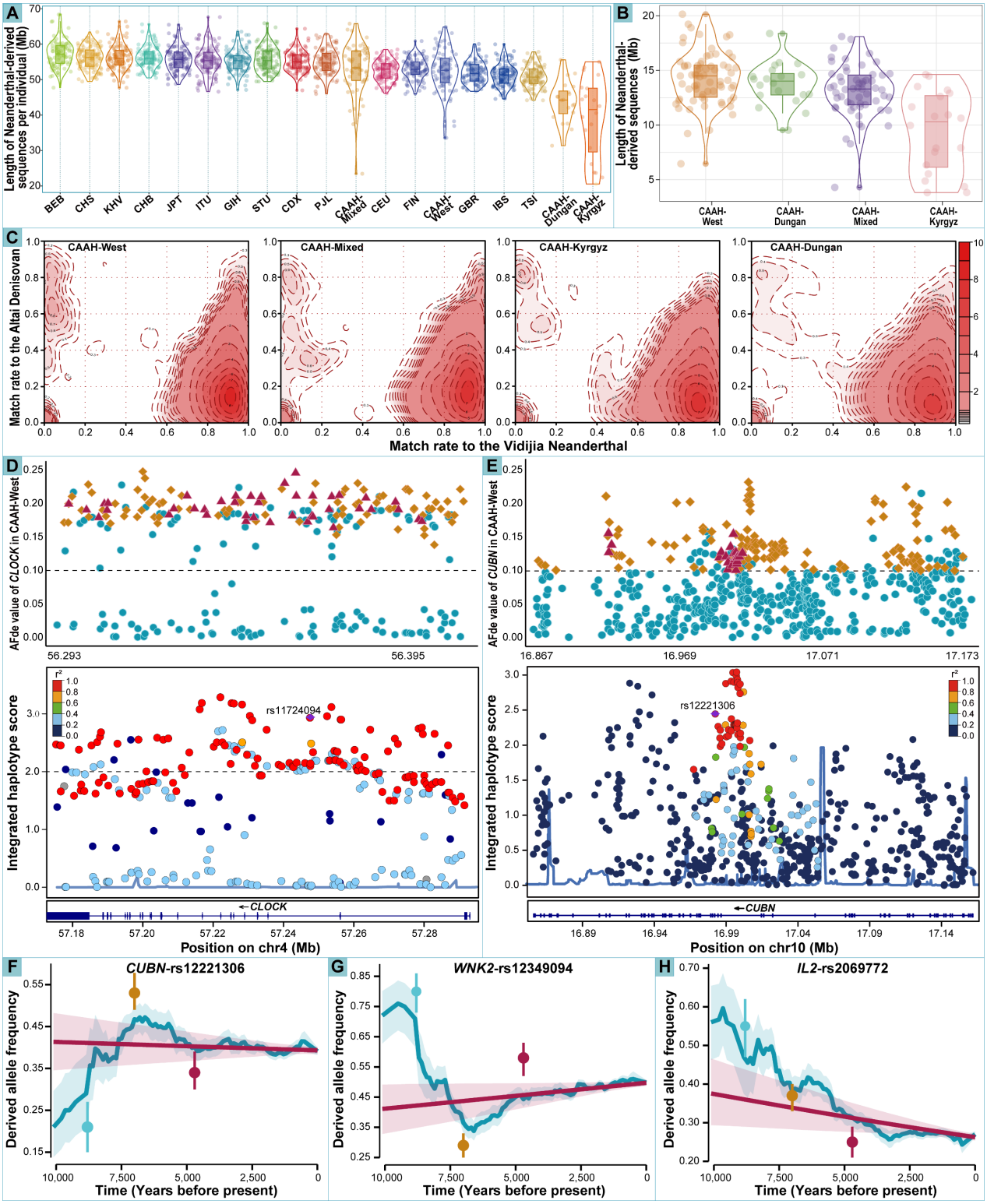


Fig. 6. Distribution of Neanderthal- and Denisovan-derived archaic introgression segments (AISs) and adaptive introgression signals. (**A**), Distribution of Neanderthal-derived AISs identified via the IBDmix method across four CAAH subgroups and Eurasian reference populations from the 1000 Genomes Project (1KGP). (**B**), Distribution of high-confidence Neanderthal-inherited AISs within each CAAH subgroup, as determined by IBDmix and Sprime. (**C**), Contour density plots of the matching proportions of introgressed fragments to the Neanderthal and Denisovan genomes. (**D**), Distribution of AFd_e and integrated haplotype score (iHS) for potential Neanderthal-introgressed variants in the *CLOCK* gene identified in CAAH-West. The upper panel shows the AFd_e distribution. Claret triangles indicate adaptive

introgression signals, orange-yellow rhombi represent natural selection signatures, and turquoise circles denote non-adaptive introgression and non-natural selection signals in the *CLOCK* gene. The lower panel depicts iHS values and linkage patterns within the target region centered on rs11724094, with variant positions corresponding to those in the upper panel. (**E**), Distribution of AFd_e and iHS values for potential Neanderthal-introgressed variants in the *CUBN* gene identified in CAAH-West, along with linkage patterns within the target region centered on rs12221306. Variant color coding is consistent with Fig. 6D. (**F–H**), DAF trajectories over time for *CUBN*-rs12221306 (**F**), *WNK2*-rs12349094 (**G**), and *IL2*-rs2069772 (**H**).

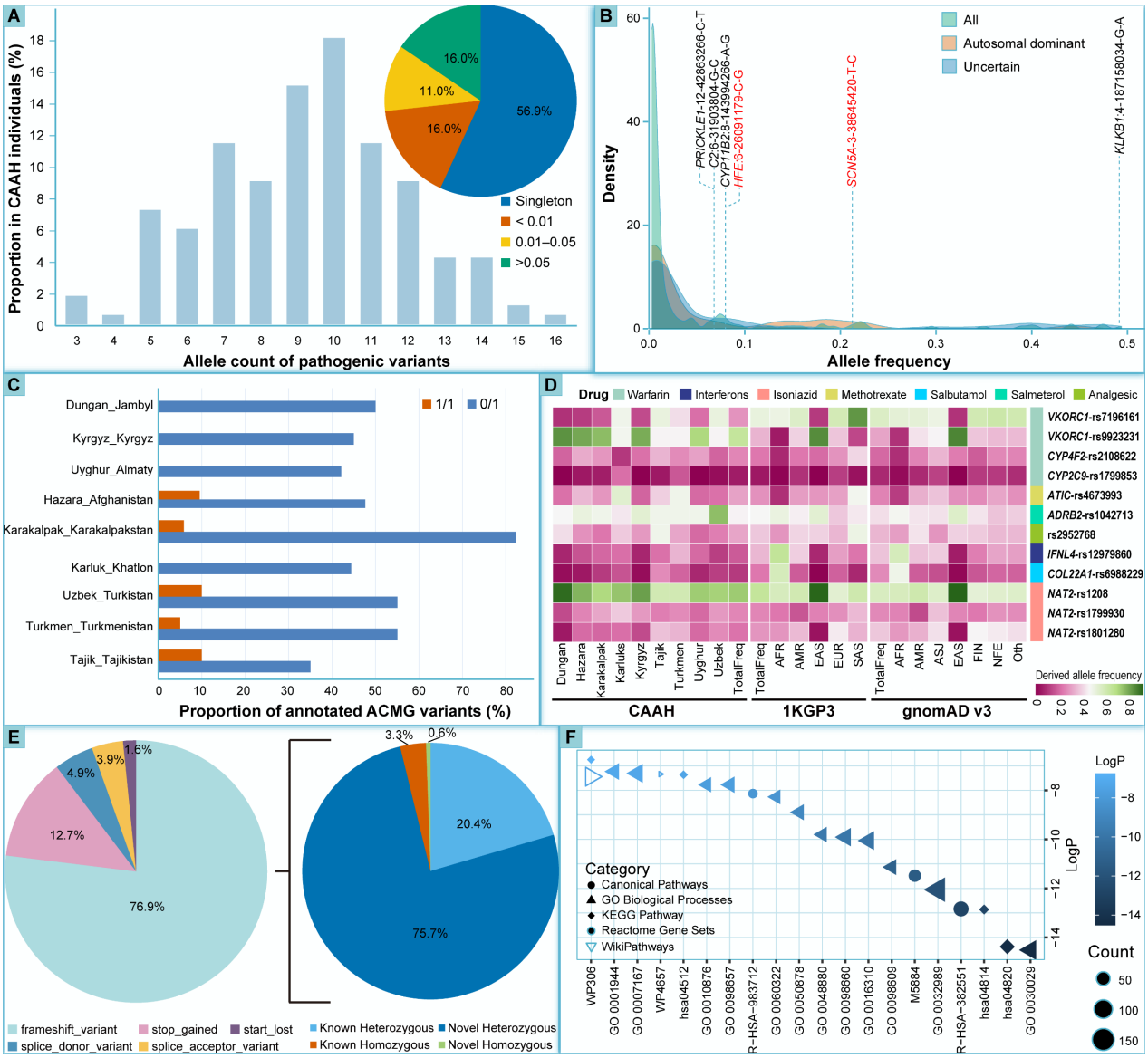
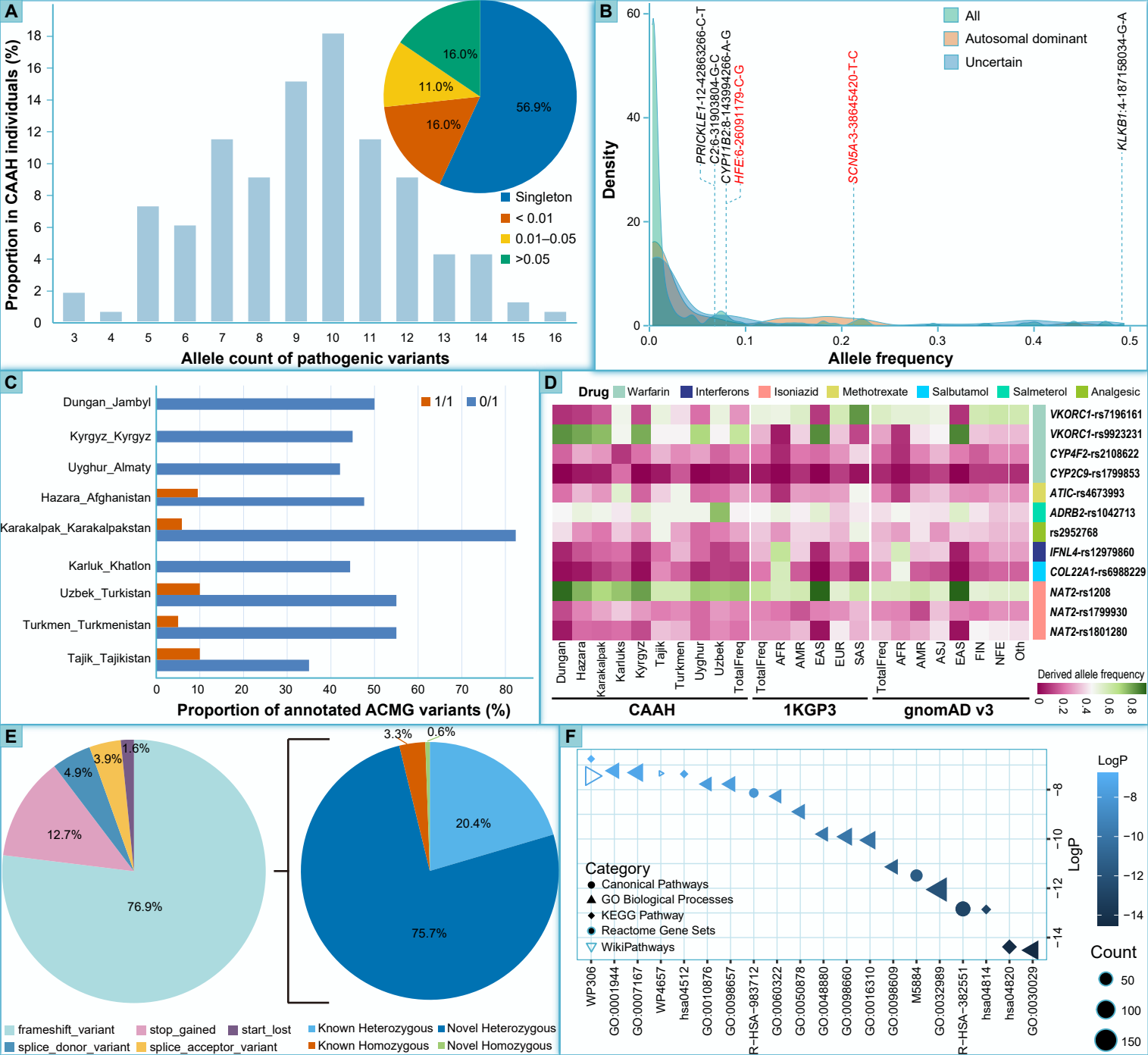
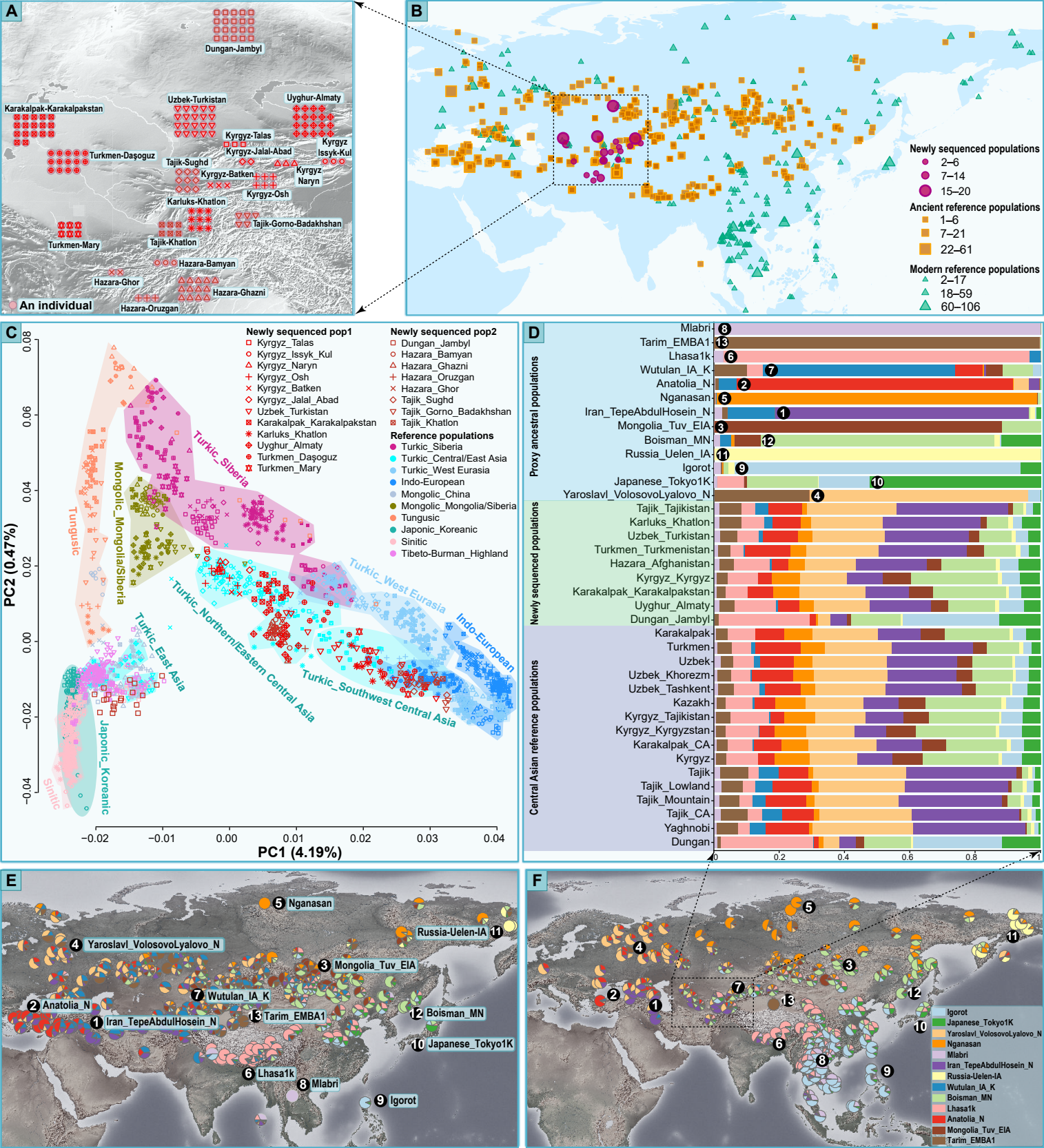
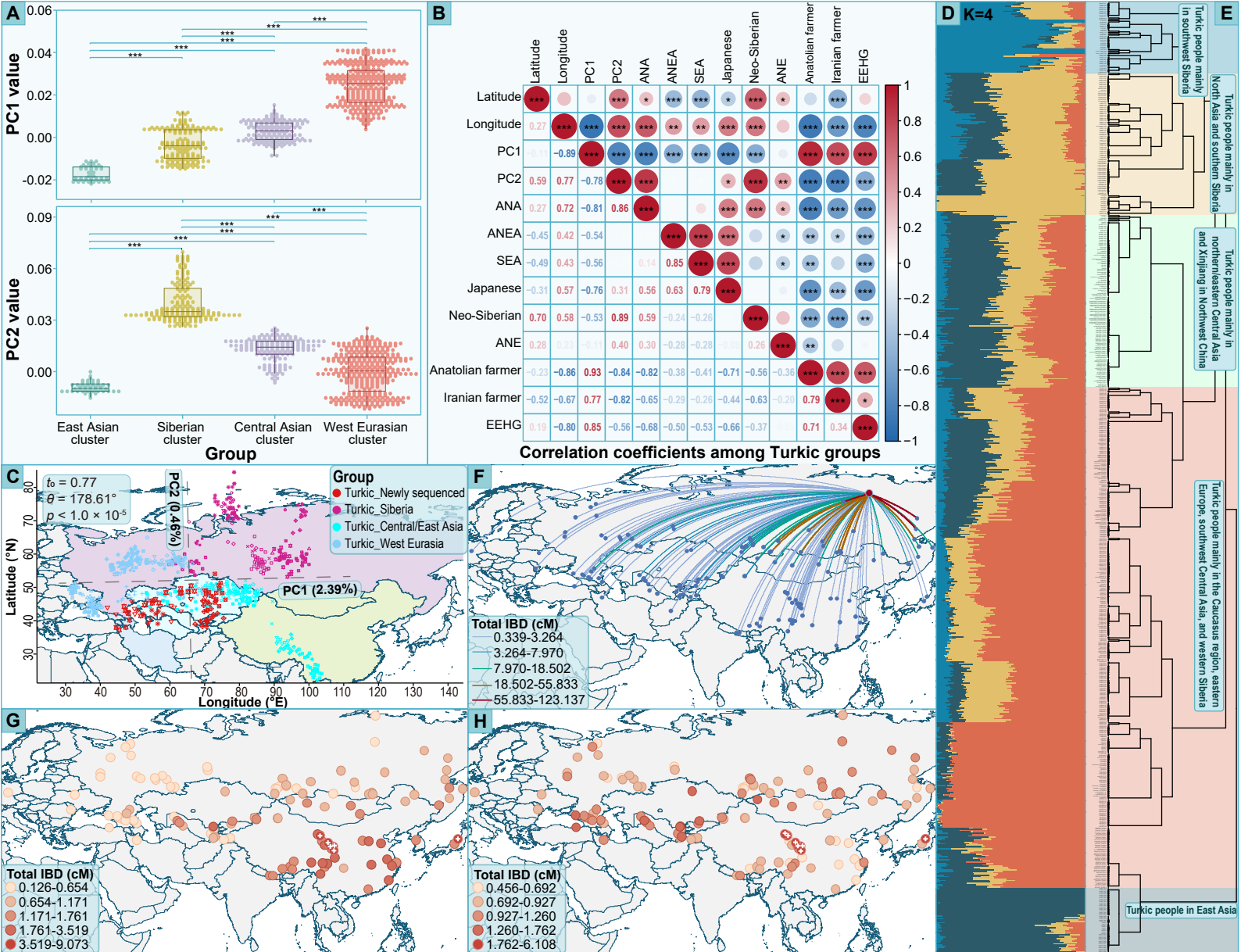


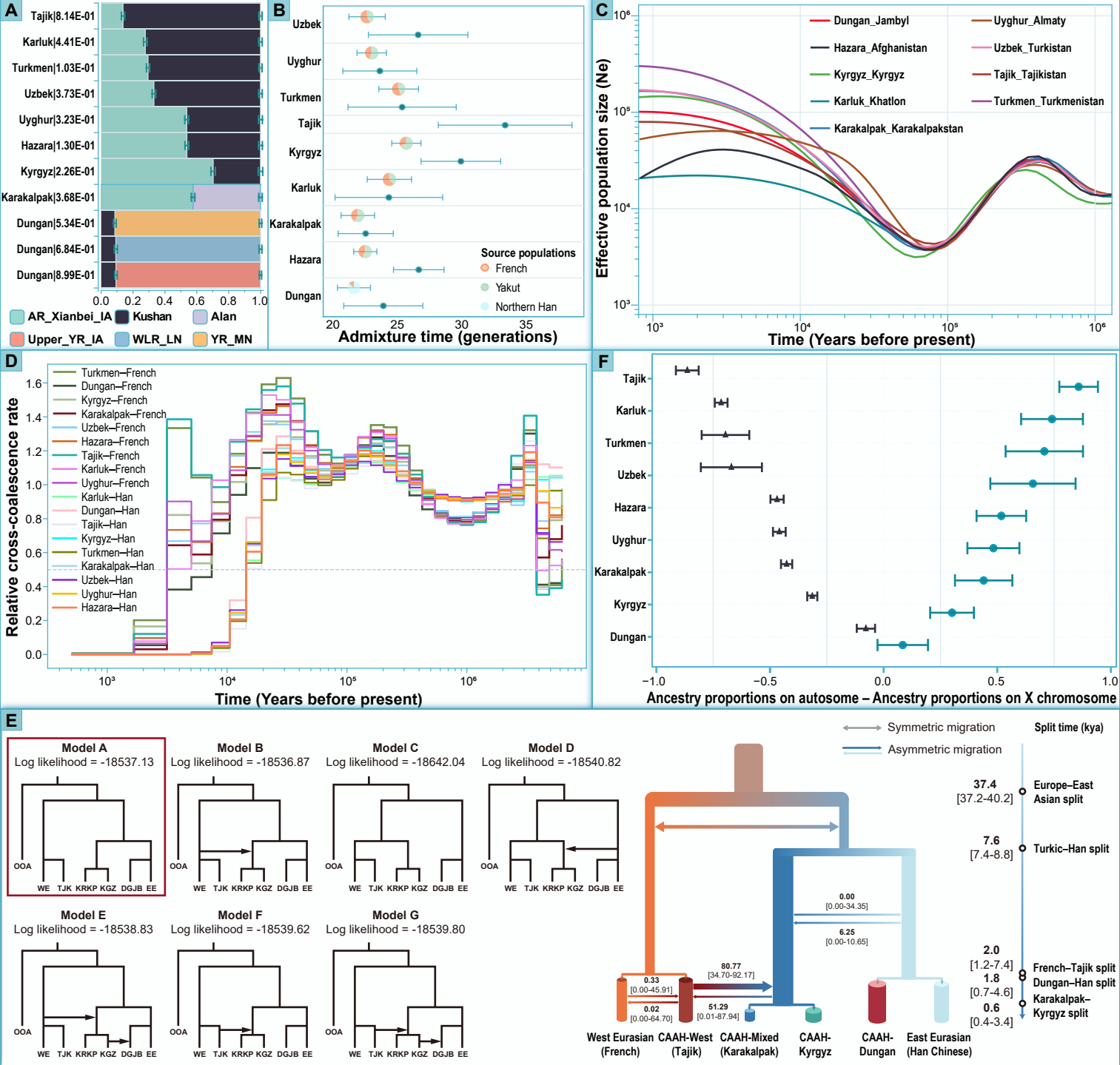
Fig. 7. Statistics of medically relevant genetic variants in our dataset. (**A**), The number of alleles of identified pathogenic variants carried by each CAAH individual, alongside the AF distribution of these variants. (**B**), Density plot displaying the frequencies of pathogenic and likely pathogenic ClinVar variants (n = 208), categorized by the most common inheritance patterns of associated monogenic disease genes; variants with autosomal dominant inheritance or unknown modes are highlighted. The variant with the highest minor allele frequency (MAF) value, three variants with MAFs > 0.05 in CAAH but average MAFs < 0.05 in the Genome Aggregation Database (gnomAD v3.1.2), and two reportable ACMG variants with MAFs > 0.05 are presented as gene names: chromosome-base pair position-reference allele-alternative allele. (**C**), Distribution of homozygous (1/1) and heterozygous (0/1) genotypes for annotated ACMG

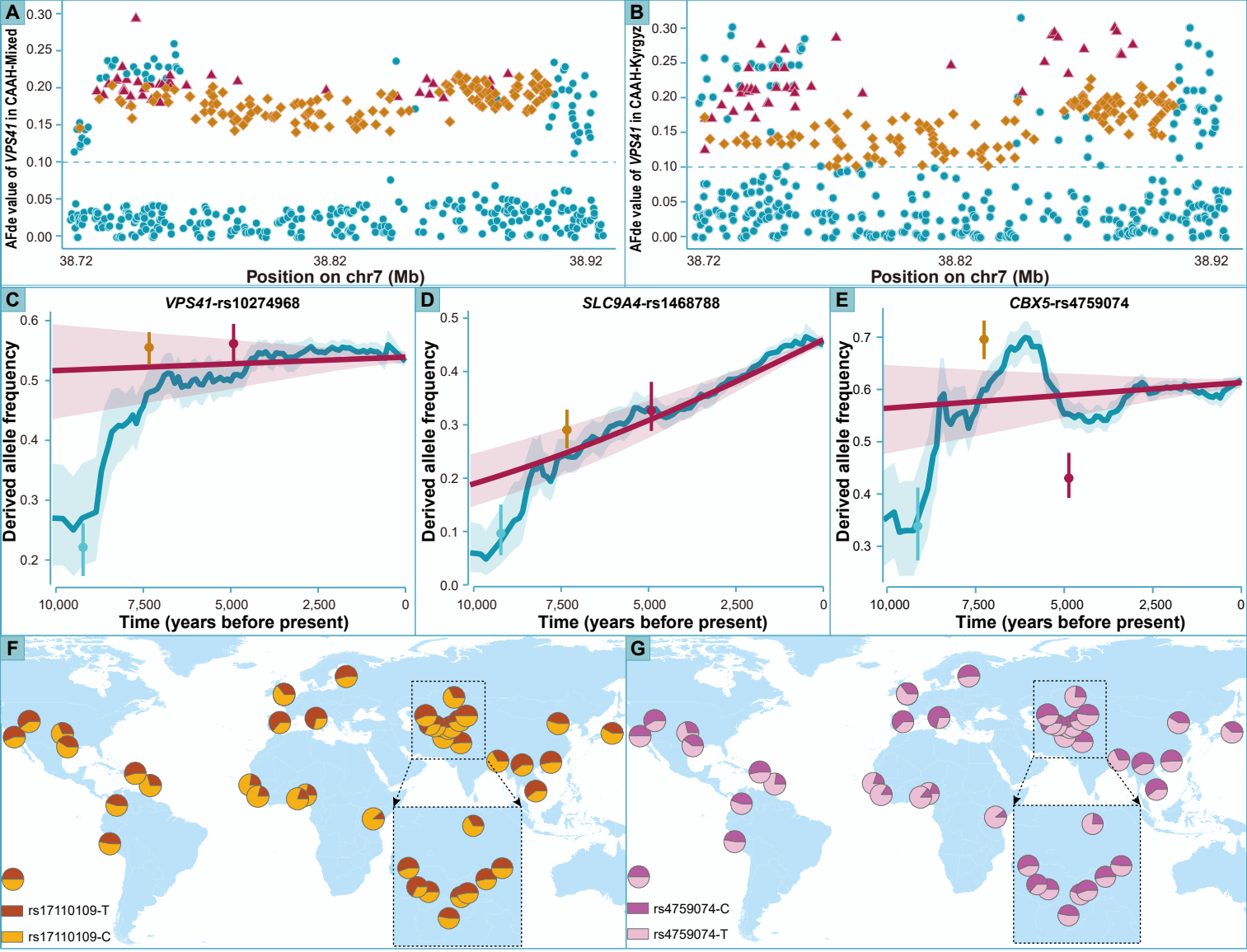
variants across each CAAH population. (**D**), Frequency spectra of derived alleles for pharmacogenomic variants in our dataset, the 1KGP, and gnomAD. AFR: African, AMR: American, EAS: East Asian, EUR: European, SAS: South Asian, ASJ: Ashkenazi Jewish, FIN: Finnish, NFE: non-Finnish European, and Oth: Remaining individuals. (**E**), Functional annotations of putative loss-of-function variants and the proportions of previously unreported and known heterozygous and homozygous protein-truncating variants (PTVs) in our dataset. (**F**), The enrichment results of annotated heterozygous PTVs.

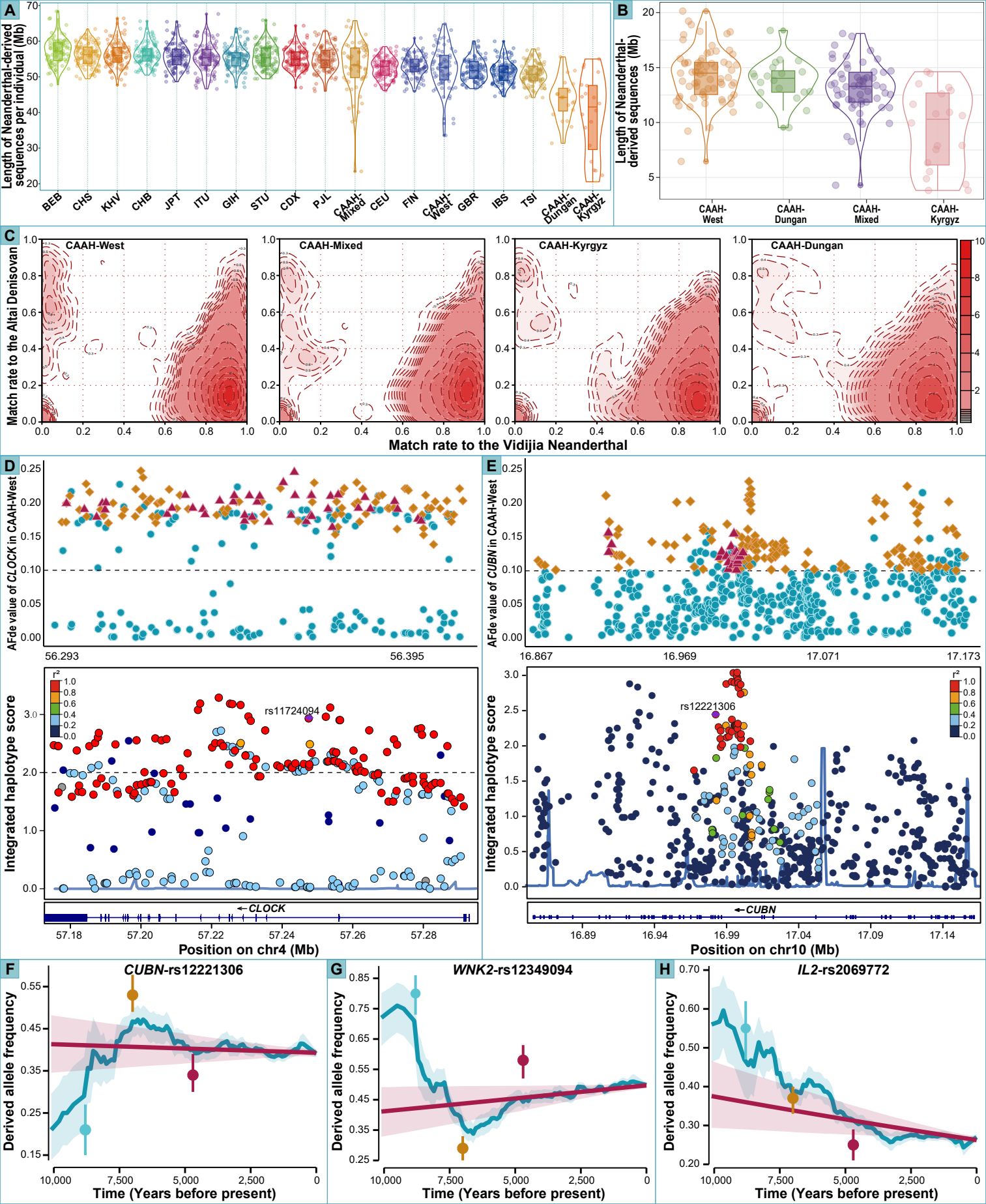

A
Proportion in CAAH individuals (%)
Allele count of pathogenic variants
56.9%
16.0%
11.0%
16.0%
Singleton
< 0.01
0.01–0.05
>0.05
B
Density
Allele frequency
All
Autosomal dominant
Uncertain
PRICKLE1-12-42863266-C-T
C2:6-31903804-G-C
CYP11B2:8-143994266-A-G
HFE:6-26091179-C-G
SCN5A-3-38645420-T-C
KLKB1:4-187158034-G-A
C
Dungan_Jambyl
Kyrgyz_Kyrgyz
Uyghur_Almaty
Hazara_Afghanistan
Karakalpak_Karakalpakstan
Karluk_Khatlon
Uzbek_Turkistan
Turkmen_Turkmenistan
Tajik_Tajikistan
1/1
0/1
Proportion of annotated ACMG variants (%)
D
Drug
Warfarin
Interferons
Isoniazid
Methotrexate
Salbutamol
Salmeterol
Analgesic
VKORC1-rs7196161
VKORC1-rs9923231
CYP4F2-rs2108622
CYP2C9-rs1799853
ATIC-rs4673993
ADRB2-rs1042713
rs2952768
IFNL4-rs12979860
COL22A1-rs6988229
NAT2-rs1208
NAT2-rs1799930
NAT2-rs1801280
Dungan
Hazara
Karakalpak
Karluks
Kyrgyz
Tajik
Turkmen
Uyghur
Uzbek
TotalFreq
AFR
AMR
EAS
EUR
SAS
ASJ
FIN
NFE
Oth
CAAH
1KGP3
gnomAD v3
Derived allele frequency
E
76.9%
12.7%
4.9%
3.9%
1.6%
3.3%
0.6%
20.4%
75.7%
frameshift_variant
stop_gained
start_lost
splice_donor_variant
splice_acceptor_variant
Known Heterozygous
Novel Heterozygous
Known Homozygous
Novel Homozygous
F
LogP
Category
Canonical Pathways
GO Biological Processes
KEGG Pathway
Reactome Gene Sets
WikiPathways
Count
WP306
GO:0001944
GO:0007167
WP4657
hsa04512
GO:0010876
GO:0098657
R-HSA-983712
GO:0060322
GO:0050878
GO:0048880
GO:0098660
GO:0016310
GO:0098609
M5884
GO:0032989
R-HSA-382551
hsa04814
hsa04820
GO:0030029


A
Dungan-Jambyl
Karakalpak-Karakalpakstan
Uzbek-Turkistan
Uyghur-Almaty
Kyrgyz-Talas
Kyrgyz-Jalal-Abad
Kyrgyz Issyk-Kul
Turkmen-Daşoguz
Tajik-Sughd
Kyrgyz-Batken
Kyrgyz Naryn
Kyrgyz-Osh
Karluks-Khatlon
Tajik-Gorno-Badakhshan
Turkmen-Mary
Tajik-Khatlon
Hazara-Bamyan
Hazara-Ghor
Hazara-Ghazni
Hazara-Oruzgan
An individual
B
Newly sequenced populations
2–6
7–14
15–20
Ancient reference populations
1–6
7–21
22–61
Modern reference populations
2–17
18–59
60–106
C
Newly sequenced pop1
Kyrgyz_Talas
Kyrgyz_Issyk_Kul
Kyrgyz_Naryn
Kyrgyz_Osh
Kyrgyz_Batken
Kyrgyz_Jalal_Abad
Uzbek_Turkistan
Karakalpak_Karakalpakstan
Karluks_Khatlon
Uyghur_Almaty
Turkmen_Daşoguz
Turkmen_Mary
Newly sequenced pop2
Dungan_Jambyl
Hazara_Bamyan
Hazara_Ghazni
Hazara_Oruzgan
Hazara_Ghor
Tajik_Sughd
Tajik_Gorno_Badakhshan
Tajik_Khatlon
Reference populations
Turkic_Siberia
Turkic_Central/East Asia
Turkic_West Eurasia
Indo-European
Mongolic_China
Mongolic_Mongolia/Siberia
Tungusic
Japonic_Koreanic
Sinitic
Tibeto-Burman_Highland
Turkic_Siberia
Mongolic_Mongolia/Siberia
Tungusic
Turkic_West Eurasia
Indo-European
Turkic_Northern/Eastern Central Asia
Turkic_East Asia
Turkic_Southwest Central Asia
Japonic_Koreanic
Sinitic
PC2 (0.47%)
PC1 (4.19%)
D
Proxy ancestral populations
Newly sequenced populations
Central Asian reference populations
Mlabri
Tarim_EMBA1
Lhasa1k
Wutulan_IA_K
Anatolia_N
Nganasan
Iran_TepeAbdulHosein_N
Mongolia_Tuv_EIA
Boisman_MN
Russia_Uelen_IA
Igorot
Japanese_Tokyo1K
Yaroslavl_VolosovoLyalovo_N
Tajik_Tajikistan
Karluks_Khatlon
Uzbek_Turkistan
Turkmen_Turkmenistan
Hazara_Afghanistan
Kyrgyz_Kyrgyz
Karakalpak_Karakalpakstan
Uyghur_Almaty
Dungan_Jambyl
Karakalpak
Turkmen
Uzbek
Uzbek_Khorezm
Uzbek_Tashkent
Kazakh
Kyrgyz_Tajikistan
Kyrgyz_Kyrgyzstan
Karakalpak_CA
Kyrgyz
Tajik
Tajik_Lowland
Tajik_Mountain
Tajik_CA
Yaghnobi
Dungan
E
Nganasan
Russia-Uelen-IA
Yaroslavl_VolosovoLyalovo_N
Mongolia_Tuv_EIA
Wutulan_IA_K
Anatolia_N
Iran_TepeAbdulHosein_N
Tarim_EMBA1
Boisman_MN
Japanese_Tokyo1K
Lhasa1k
Mlabri
Igorot
F
Igorot
Japanese_Tokyo1K
Yaroslavl_VolosovoLyalovo_N
Nganasan
Mlabri
Iran_TepeAbdulHosein_N
Russia-Uelen-IA
Wutulan_IA_K
Boisman_MN
Lhasa1k
Anatolia_N
Mongolia_Tuv_EIA
Tarim_EMBA1


A
PC1 value
PC2 value
East Asian cluster
Siberian cluster
Central Asian cluster
West Eurasian cluster
Group
B
Latitude
Longitude
PC1
PC2
ANA
ANEA
SEA
Japanese
Neo-Siberian
ANE
Anatolian farmer
Iranian farmer
EEHG
Correlation coefficients among Turkic groups
C
t_0 = 0.77
θ = 178.61°
p < 1.0 × 10^{-5}
PC2 (0.46%)
PC1 (2.39%)
Latitude (°N)
Longitude (°E)
Group
Turkic_Newly sequenced
Turkic_Siberia
Turkic_Central/East Asia
Turkic_West Eurasia
D
K=4
E
Turkic people mainly in southwest Siberia
Turkic people mainly in North Asia and southern Siberia
Turkic people mainly in northeastern Central Asia and Xinjiang in Northwest China
Turkic people mainly in the Caucasus region, eastern Europe, southwest Central Asia, and western Siberia
Turkic people in East Asia
F
Total IBD (cM)
0.339-3.264
3.264-7.970
7.970-18.502
18.502-55.833
55.833-123.137
G
Total IBD (cM)
0.126-0.654
0.654-1.171
1.171-1.761
1.761-3.519
3.519-9.073
H
Total IBD (cM)
0.456-0.692
0.692-0.927
0.927-1.260
1.260-1.762
1.762-6.108


A
Tajik|8.14E-01
Karluk|4.41E-01
Turkmen|1.03E-01
Uzbek|3.73E-01
Uyghur|3.23E-01
Hazara|1.30E-01
Kyrgyz|2.26E-01
Karakalpak|3.68E-01
Dungan|5.34E-01
Dungan|6.84E-01
Dungan|8.99E-01
0.0 0.2 0.4 0.6 0.8 1.0
AR_Xianbei_IA
Kushan
Alan
Upper_YR_IA
WLR_LN
YR_MN
B
Uzbek
Uyghur
Turkmen
Tajik
Kyrgyz
Karluk
Karakalpak
Hazara
Dungan
Source populations
French
Yakut
Northern Han
20 25 30 35
Admixture time (generations)
C
Dungan_Jambyl
Hazara_Afghanistan
Kyrgyz_Kyrgyz
Karluk_Khatlon
Karakalpak_Karakalpakstan
Uyghur_Almaty
Uzbek_Turkistan
Tajik_Tajikistan
Turkmen_Turkmenistan
Effective population size (Ne)
Time (Years before present)
D
Turkmen–French
Dungan–French
Kyrgyz–French
Karakalpak–French
Uzbek–French
Hazara–French
Tajik–French
Karluk–French
Uyghur–French
Karluk–Han
Dungan–Han
Tajik–Han
Kyrgyz–Han
Turkmen–Han
Karakalpak–Han
Uzbek–Han
Uyghur–Han
Hazara–Han
Relative cross-coalescence rate
Time (Years before present)
F
Tajik
Karluk
Turkmen
Uzbek
Hazara
Uyghur
Karakalpak
Kyrgyz
Dungan
−1.0 −0.5 0.0 0.5 1.0
Ancestry proportions on autosome – Ancestry proportions on X chromosome
E
Model A
Log likelihood = -18537.13
Model B
Log likelihood = -18536.87
Model C
Log likelihood = -18642.04
Model D
Log likelihood = -18540.82
Model E
Log likelihood = -18538.83
Model F
Log likelihood = -18539.62
Model G
Log likelihood = -18539.80
OOA
WE TJK KRKP KGZ DGJB EE
Symmetric migration
Asymmetric migration
Split time (kya)
37.4 [37.2-40.2] Europe–East Asian split
7.6 [7.4-8.8] Turkic–Han split
2.0 [1.2-7.4] French–Tajik split
1.8 [0.7-4.6] Dungan–Han split
0.6 [0.4-3.4] Karakalpak–Kyrgyz split
0.00 [0.00-34.35]
6.25 [0.00-10.65]
80.77 [34.70-92.17]
51.29 [0.01-87.94]
0.33 [0.00-45.91]
0.02 [0.00-64.70]
West Eurasian (French)
CAAH-West (Tajik)
CAAH-Mixed (Karakalpak)
CAAH-Kyrgyz
CAAH-Dungan
East Eurasian (Han Chinese)


A
AFde value of VPS41 in CAAH-Mixed
0.30
0.25
0.20
0.15
0.10
0.05
0.00
38.72
38.82
38.92
Position on chr7 (Mb)
B
AFde value of VPS41 in CAAH-Kyrgyz
0.30
0.25
0.20
0.15
0.10
0.05
0.00
38.72
38.82
38.92
Position on chr7 (Mb)
C
VPS41-rs10274968
Derived allele frequency
0.6
0.5
0.4
0.3
0.2
10,000
7,500
5,000
2,500
0
Time (years before present)
D
SLC9A4-rs1468788
Derived allele frequency
0.5
0.4
0.3
0.2
0.1
0.0
10,000
7,500
5,000
2,500
0
Time (years before present)
E
CBX5-rs4759074
Derived allele frequency
0.7
0.6
0.5
0.4
0.3
10,000
7,500
5,000
2,500
0
Time (years before present)
F
rs17110109-T
rs17110109-C
G
rs4759074-C
rs4759074-T


A
Length of Neanderthal-derived sequences per individual (Mb)
BEB
CHS
KHV
CHB
JPT
ITU
GIH
STU
CDX
PJL
CAAH-Mixed
CEU
FIN
CAAH-West
GBR
IBS
TSI
CAAH-Dungan
CAAH-Kyrgyz
B
Length of Neanderthal-derived sequences (Mb)
CAAH-West
CAAH-Dungan
CAAH-Mixed
CAAH-Kyrgyz
C
CAAH-West
CAAH-Mixed
CAAH-Kyrgyz
CAAH-Dungan
Match rate to the Altai Denisovan
Match rate to the Vidijia Neanderthal
D
AFde value of CLOCK in CAAH-West
rs11724094
Integrated haplotype score
CLOCK
Position on chr4 (Mb)
E
AFde value of CUBN in CAAH-West
rs12221306
Integrated haplotype score
CUBN
Position on chr10 (Mb)
F
CUBN-rs12221306
Derived allele frequency
Time (Years before present)
G
WNK2-rs12349094
Derived allele frequency
Time (Years before present)
H
IL2-rs2069772
Derived allele frequency
Time (Years before present)