Ancient farmer and steppe pastoralist-related founding lineages contributed to the complex landscape of episodes in the diversification of Chinese paternal lineages

Mengge Wang[1,2,3,*,#], Yuguo Huang[1*], Kaijun Liu[4,5*], Haibing Yuan[2], Shuhan Duan[1,6], Zhiyong Wang[1,7], Lanhai Wei[8], Hongbing Yao[9], Qiuxia Sun[1,10], Jie Zhong[1], Renkuan Tang[10], Jing Chen[1,11], Yuntao Sun[1,12], Xiangping Li[1,7], Haoran Su[1,14], Qingxin Yang[7], Liping Hu[7], Libing Yun[12], Junbao Yang[13], Shengjie Nie[7], Yan Cai[14], Jiangwei Yan[11], Kun Zhou[5], 10K_CPGDP Consortium†, Chuanchao Wang[15], Bofeng Zhu[16,17,#], Chao Liu[16,18,#], Guanglin He[1,2,*,#]

[1]Institute of Rare Diseases, West China Hospital of Sichuan University, Sichuan University, Chengdu, 610000, China
[2]Center for Archaeological Science, Sichuan University, Chengdu, 610000, China
[3]Faculty of Forensic Medicine, Zhongshan School of Medicine, Sun Yat-sen University, Guangzhou, 510275, China
[4]School of International Tourism and culture, Guizhou Normal University, Guiyang, 550025, China
[5]Chengdu 23Mofang Biotechnology Co., Ltd., Tianfu Software Park, Chengdu, Sichuan 610042, China
[6]School of Basic Medical Sciences, North Sichuan Medical College, Nanchong, 637100, China
[7]School of Forensic Medicine, Kunming Medical University, Kunming, 650500, China
[8]School of Ethnology and Anthropology, Institute of Humanities and Human Sciences, Inner Mongolia Normal University, Hohhot, 010022, China
[9]Belt and Road Research Center for Forensic Molecular Anthropology Gansu University of Political Science and Law, Lanzhou, 730000, China
[10]Department of Forensic Medicine, College of Basic Medicine, Chongqing Medical University, Chongqing, 400331, China
[11]School of Forensic Medicine, Shanxi Medical University, Jinzhong, 030001, China
[12]Institute of Forensic Medicine, West China School of Basic Medical Sciences & Forensic Medicine, Sichuan University, Chengdu, 610041, China
[13]Institute of Basic Medicine and Forensic Medicine, North Sichuan Medical College and Center for Genetics and Prenatal diagnosis, Affiliated Hospital of North Sichuan Medical College, Nanchong, Sichuan, 637007, China
[14]Department of Clinical Laboratory, North Sichuan Medical College and Center for Genetics and Prenatal diagnosis, Affiliated Hospital of North Sichuan Medical College, Nanchong, Sichuan, 637007, China
[15]State Key Laboratory of Cellular Stress Biology, School of Life Sciences, Xiamen University, Xiamen, 361005, China
[16]Guangzhou Key Laboratory of Forensic Multi-Omics for Precision Identification, School of Forensic Medicine, Southern Medical University, Guangzhou, 510515, China
[17]Microbiome Medicine Center, Department of Laboratory Medicine, Zhujiang Hospital, Southern Medical University, Guangzhou, Guangdong, 510515, China
[18]Anti-Drug Technology Center of Guangdong Province, Guangzhou, 510230, China

†Full author list from the 10K_CPGDP Consortium was presented at the end of the manuscript

*These authors contributed equally to this work
[#]**Corresponding authors:** E-mails: Menggewang2021@163.com, zhubofeng7372@126.com liuchaogzf@163.com, guanglinhescu@163.com

Abstract (250)

Ancient DNA advances have reported the complex genetic history of Eurasians, but how the knowledge of ancient subsistence strategy shifts and population movements influenced the fine-scale paternal genetic structure in East Asia has not been assessed. Here, we reported one integrated Y-chromosome genomic database of 15,530 people, including 1753 ancient people and newly-reported 919 individuals genotyped using our recently-developed targeted sequencing YHSeqY3000 panel, to explore Chinese genomic diversity, population evolutionary tracts and their genetic formation mechanism. We identified four major ancient technological innovations and population movements that shaped the landscape of Chinese paternal lineages. First, the expansion of millet farmers and early East Asians from the Yellow River Basin carrying the major O2/D subclades promoted the formation of the Sino-Tibetan people's major composition and accelerated the Tibetan Plateau's permanent occupation. Second, rice farmers' dispersal from the Yangtze River Valley carrying O1 and some sublineages of O2 contributed significantly to Tai-Kadai, Austronesian, Hmong-Mien, Austroasiatic people and southern Han Chinese. Third, Siberian-related paternal lineages of Q and C originated and boomed from Neolithic hunter-gatherers from the Mongolian Plateau and the Amur River Basin and significantly influenced the gene pools of northern Chinese. Fourth, western Eurasian-derived J, G and R lineages initially spread with Yamnaya steppe pastoralists and other proto-Indo-European people and further widely dispersed via the trans-Eurasian cultural communication along the Eurasian Steppe and the ancient Silk Road, remaining genetic trajectories in northwestern Chinese. Our work provided comprehensive modern and ancient genetic evidence to illuminate the impact of population interaction from the ancient farmer or herder-based societies on the genetic diversity patterns of modern people, revised our understandings of ancestral sources of Chinese paternal lineages, underscored the scientific imperative of the large-scale genomic resources of dense spatiotemporal underrepresented sampling populations to understand human evolutionary history.

Keywords: Evolutionary history, Y-chromosome phylogeny, YanHuang cohort, Founding lineage

Introduction

Comprehensively documenting the genetic landscape of genetically diverse worldwide populations and illuminating their influence on the human demographical history and genetic basis of complex traits and diseases was the goal of population genomics and human pangenome projects (Bergstrom et al. 2020; Byrska-Bishop et al. 2022). Current genetic resources, especially large-scale population genomic cohorts, were mainly derived from descendants of European ancestry, which limited the transferability of European-based genetic findings in other non-European people as their differences in the allele frequency spectrum, linkage disequilibrium, effect size and different trajectories of the evolutionary process (Sirugo et al. 2019). Previous anthropological and genetic findings suggested that the genetically underrepresented African continent possessing the most linguistic (Afro-Asiatic, Nilo-Saharan, Niger-Congo and KhoeSan) and genetic diversity was regarded as the cradle of modern human origin (Choudhury et al. 2020). Similarly, East Asia, including countries of China, Mongolia, Japan, North Korea and South Korea, served as one of the earliest cradles of civilization and the crossroad of the peopling of Oceania, Siberia and America, whose genetic landscape is also poorly characterized in the era of population-based genomics. China is the world's second-most populous country with a population size exceeding 1.4 billion, and there are nearly 300 living languages that belong to seven language families [Sino-Tibetan (ST), Altaic (Mongolic, Tungusic, Turkic, Japonic and Koreanic), Tai-Kadai (TK), Hmong-Mien (HM), Austronesian (AN), Austroasiatic (AA) and Indo-European (IE)] (Lewis et al. 2016). China harbors substantial genetic, physical, cultural and ethnolinguistic diversity, which allows this region to maintain an unrivaled position in the study of the complex demographic history of ethnolinguistically diverse populations, including human divergence, migration and admixture events, as well as interrelationships between genetics and cultures (Wang, Yeh, et al. 2021; Kumar et al. 2022; Zhang et al. 2022). It has been well-documented that European bias can cause human health inequality and non-transferability of PRS across genetically different populations (Sirugo et al. 2019). The worldwide genetic consortiums have conducted many studies or projects to fill the gap in the missing genetic diversity in human genetics and genomics and dissect the genetic basis of complex traits/diseases and the evolutionary history of Chinese populations. Recently, a large body of research focused on geographically different ST, Mongolic, Tungusic, Turkic, TK and HM groups has been carried out using genome-wide SNP arrays (Feng et al. 2017; He, Wang, Li, et al. 2021; Ma et al. 2021; He et

al. 2022; Wang, He, et al. 2022; Wang et al. 2023). Moreover, there has been a rapid increase of whole-genome sequencing (WGS) studies on ethnolinguistically diverse Chinese populations, such as the Westlake BioBank for Chinese (WBBC), NyuWa genome resource, China Metabolic Analytics Project (ChinaMAP), 10K Chinese People Genomic Diversity Project (10K_CPGDP) and STROMICS (Cao et al. 2020; Zhang, Luo, et al. 2021; Cong et al. 2022; Cheng et al. 2023; He, Yao, et al. 2023). Generally, these studies advanced our understanding of demographic history, the patterns of genetic diversity and the genetic architecture of complex traits/diseases in Chinese populations from autosome-related perspectives. More unknown genetic features from a uniparental perspective and population-scale project should be further explored.

A complex assortment of neutral and non-neutral selection processes has shaped the genetic landscape of present-day humans. Autosome-based genomic studies could provide basal insights into the population demographic and adaptive history, and mitochondrial and Y-chromosomal variants could provide additional signals of evolutionary processes that have uniquely left signatures in male/female specific regions in the Y/mitochondrial DNA (mtDNA) genome (Poznik et al. 2016; Nielsen et al. 2017; Li et al. 2019). The non-recombining part of the Y-chromosome (NRY) is strictly paternally inherited, transmitted through the germ line and not affected by heteroplasmy (Jobling and Tyler-Smith 2017), namely unique features of haploidy, escape from crossing over and male specificity. The length of the Y-chromosome is approximately 4,000 times larger than that of mtDNA. Therefore, it contains much more information content relative to mtDNA. The mutation rates of Y-chromosomal single nucleotide polymorphisms (Y-SNPs) are lower than that of mtDNA and short tandem repeats (STRs), and Y-chromosomal variants show more obvious geographical specificity than other markers. These properties have led to its genetic variations becoming a significant part of studies of human evolutionary history on different time scales (Poznik et al. 2016; Jobling and Tyler-Smith 2017).

Improved sequencing technologies and computational innovations in genome assembly, read mapping, variant calling and benchmarking have promoted the generation of many complete Y-chromosomal sequences and greatly enriched our understanding of NRY variations over the past few years (Olson et al. 2023). These identified Y-chromosomal variants enable a robust phylogenetic tree to be gradually constructed, in which the branch lengths are proportional to the number of mutations (Poznik et al. 2016; Jobling and Tyler-Smith 2017; Zhabagin et al. 2022). In addition to constructing a high-quality and high-resolution phylogeny with a robust topology structure and divergence models, several broad approaches have been applied to estimate the mutation rates of Y-SNPs, which could facilitate the conversion of branch length information into time and enable the estimation of the time to the most recent common ancestor (TMRCA) shared by two individuals (Xue et al. 2009; Francalacci et al. 2013; Poznik et al. 2013; Helgason et al. 2015; Jobling and Tyler-Smith 2017).

Many vital genetic works have been conducted to trace one's ancestor through the paternal lineages and have been an essential source of phylogenetic information for studies of human origin, migration and admixture history in the past two decades of the human genome era (Jobling and Tyler-Smith 2017). Su et al. genotyped 19 Y-SNPs in 925 geographically and genetically different male samples and identified a tremendous northward migration into East Asia during the last Ice Age (Su et al. 1999). Soon after, Ke et al. genotyped three Y-chromosomal biallelic markers in 12,127 male individuals from 163 populations and confirmed the Out-of-Africa hypothesis (Ke et al. 2001). Y-chromosomes can also document trajectories of recent population migration, admixture and expansion events. Zerjal et al. investigated the genetic diversity of 2,123 males using genotype data of 32 Y-chromosomal markers, and the observed pattern of the star-cluster phylogeny indicated the significant influence of the Mongol Empire's westward expansion on the genetic architecture of Asian populations (Zerjal et al. 2003). Large-scale genetic studies focused on the fine-scale paternal demographic histories of East Asians have been conducted over the past two decades, revealing the patterns of admixture and microevolution during the initial human settlement and subsequent migrations in East Asia (Wang and Li 2013). Resequencing whole Y-chromosomes based on next-generation sequencing (NGS) and computational technology has revolutionized the research paradigm. Wei et al. identified 6,662 high-confidence variants in 36 diverse Y-chromosomes and calibrated previously available Y-chromosomal phylogenies (Wei et al. 2013). Poznik et al. reported 1,244 complete Y-chromosomal genomes randomly sampled from 26 worldwide populations by the 1000 Genomes Project (1KGP), and they discovered more than 65,000 variants and found bursts of expansion within specific paternal lineages occurred in the last few thousand years (Poznik et al. 2016). Several Y-chromosomal investigations have also been performed on a single population or specific lineage. Wang et al. analyzed 285 Y-chromosomal sequences and identified two Neolithic

expansions of Tibeto-Burman (TB) groups and their specific paternal lineages (Wang et al. 2018). The O1a-M119 shared by Sinitic, TK and AN people, the founding paternal lineages of Tungusic or Mongolic groups (C2a-F5484 or C2b1a1a1a-M407), C2b-F1067 dominated in eastern Eurasian populations, Q1a1a-M120 connecting East Asian and Siberian populations and other specific lineages have been studied to explore their origin, diffusion and contribution to the gene pool of ethnolinguistically diverse East Asian groups (Huang, Wei, et al. 2018; Sun et al. 2019; Wu et al. 2020; Liu, Ma, et al. 2021; Sun et al. 2021). However, the large-scale genomic database was limited for China, which can provide one vital clue for exploring the entire genetic landscape of Chinese people and their ancient influence factors.

To promote the effective screening of East Asian-specific phylogenetically informative markers and develop a high-resolution Y-SNP panel for large-scale population genotyping of ethnolinguistically diverse groups, we initiated the 10K_CPGDP, one of the aims of which is to comprehensively capture the entire genetic landscape of genetically and ethnolinguistically diverse Chinese populations, including the elucidation of Y-chromosomal genetic variations and the paternal demographic history of underrepresented ethnic groups included in the YanHuang Cohort (YHC) genomic resource (He, Yao, et al. 2023). The YHC in the 10K_CPGDP was focused on presenting one high-quality population-specific Y-chromosome database, drawing one time-stamped higher-resolution Y-phylogeny and developing multiple NGS panels for medical and forensic applications via the combination of WGS and third-generation sequencing (TGS) techniques, including SNPs, STRs, InDels and other variations. We have developed the highest resolution chrY-specific targeted resequencing panel, the 'YHSeqY3000', and designed the SNP composition based on the whole-genome sequences and genome-wide SNP variations of Y-chromosomes in the YHC genomic resource. We genotyped 3002 panel-related Y-SNPs in 919 male individuals from 57 ethnolinguistically diverse Chinese populations and reported the cohort design and fine-scale paternal evolutionary history of Chinese minority ethnic groups. We presented one integrated Y-chromosome database including 15,530 individuals from modern and ancient Eurasian populations to eliminate the impact of ancient population migration, admixture and agricultural innovations on the landscape of the genetic structure of East Asians. We presented the haplogroup frequency spectrum (HFS) of modern and ancient Eurasians and identified multiple steppe pastoralist-related and agriculture-related founding lineages that formed the mosaic diversity of ethnolinguistically diverse Chinese people. The YHSeqY3000 can serve as a unique tool in the interdisciplinary research of evolutionary, population, medical and forensic genetics, such as paternal demographic history reconstruction, patrilineal biogeographic ancestry inference and forensic pedigree searching.

Results and Discussion

Genetic diversity of YanHuang cohort paternal lineages inferred from the highest resolution chrY-specific targeted resequencing YHSeqY3000 panel

YanHuang Cohort genomic dataset (**Fig. 1A**) in the 10K_CPGDP focused on the genetic diversity of male-specific regions of the Y-chromosome, which was the best resource to gain new insights into the full landscape of the paternal evolutionary history of East Asia. To characterize the genetic patterns of paternal lineages of ethnolinguistically distinct Chinese populations at a fine scale, we developed the highest resolution YHSeqY3000 panel based on the newly-updated database of Y-chromosomal sequences from the 10K_CPGDP genomic resource, including newly-identified Y-SNPs not presented in the latest ISOGG Y-DNA and Yfull phylogenetic trees (submitted), and sequenced 919 participants from 57 populations of 39 ethnic minorities using our newly-developed panel (**Fig. 1A and Table S1**). The YHSeqY3000 panel we presented here allowed simultaneous genotyping of more than 3000 Y-SNPs, covering the overwhelming majority of subclades of dominant paternal lineages of Chinese populations. We adopted three haplogroup classification methods, including Y-LineageTracker and HaploGrouper based on ISOGG2019-tree and in-house script based on the newly-reconstructed phylogenetic tree, to perform haplogroup inference simultaneously for reliable classification results. There were forty discrepancies (classified into entirely different macro-haplogroups) between the haplogroup results obtained through Y-LineageTracker and our in-house scripts, mainly including subclades of C-M130, J-M304, N-M231, O-M175, Q-M242, R-M207 and T-M184. However, there were only four discrepancies between the haplogroup results obtained through HaploGrouper and our in-house script; the samples belonging to F-M89 were classified into CF by HaploGrouper. The fundamental analysis for the research on paternal demographic history is the NRY haplogroup inference. Several tools have been developed to support the function of haplogroup classification: Y-LineageTracker (Chen et al. 2021), HaploGrouper (Jagadeesan et al. 2021), Yleaf (Ralf et al.

2018), AMY-tree (Van Geystelen et al. 2013) and cleantree (Ralf et al. 2015), but the NRY haplogroups assigned by some tools were outdated without reference to the latest phylogenetic tree. With the accumulation of Y-chromosome sequencing data, more and more novel NRY variants were identified. However, they have not been located in the up-to-date phylogenetic trees, which also led to errors or inaccuracies in haplogroup inference based on the non-updated phylogenetic topologies. Hence, it is necessary to calibrate and refine the Y-chromosome phylogenetic trees based on newly-identified Y-SNPs and build the continuously updated phylogeny into haplogroup classification tools. The new version of phylogenetic topology to be reported in the 10K_CPGDP will fill this gap. We observed 564 distinct paternal lineages that mainly fell into haplogroups C-M130, N-M231, O-M175 and R-M207, respectively, sampled from 150, 65, 583 and 46 individuals (**Figs. 1B-C and S1, Table S2**)). We found that 384 subhaplogroups belonging to C-M130, D-CTS3946, G-M201, J-M304, N-M231, O-M175, Q-M242 and R-M207 were observed only once (**Figs. 1C and S1, Table S2**). Significantly, we observed that subclade E1b1b1a-L539 of E-M96, which was found at high frequencies in East Africa, was present in eight individuals from Mongolian, Manchu and Hui populations (**Fig. 1C**). Haplogroup H-L901, one of the most dominant paternal lineages amongst populations in South Asia, was observed in ethnic minority groups in Northwest China. In addition, West Asian-derived T-M184 was observed in Turkic-speaking populations. The patterns of lineage distribution from one geographical region showed that different founding populations contributed to the Chinese paternal gene pool, mainly possibly from ancient migrations from Chinese indigenous rice or millet farmers or western Eurasian steppe people (**Fig. 1B**). We found that haplotype diversity (HD) values reached 1 in all populations with a sample size larger than 30, and haplogroup diversity (H) values ranged from 0.9537 (Zhuang) to 0.9979 (Manchu). Our observations demonstrated that the resolution and coverage of the newly-designed YHSeqY3000 panel are currently the highest, and this panel can be applied for finer haplogroup classification of Chinese populations than previously developed systems (Wang et al. 2019; Liu et al. 2022; He, Wang, et al. 2023; Tao et al. 2023).

Genetic connections and population stratification among modern and ancient Eurasians

We explored the genetic relationships and population differentiation among 13,777 modern and 1753 ancient Eurasian individuals based on the clustering patterns in the PCA (principal component analysis), MDS (multidimensional scaling analysis) and other population genetic analyses. PCA patterns in the context of Eurasians distinguished ancient western Eurasians from other East Asians (**Fig. S2A-H**). Both harbored different patterns of the dominant Y-lineages and different clustering branches on the phylogenetic tree (**Fig. S2I-J**). The clustering patterns of modern populations were consistent with the geographical divisions and linguistic affinity (**Fig. 2A**). We observed apparent population stratification between northern and southern East Asians and fine-scale substructure among geographically different but linguistically similar groups (**Figs. 2B-D and S3A-B**). Most AN/TB-speaking people were separated from others, but other populations possessed much-overlapped clustering positions. The ancient population from the Iron Age Hanben site was clustered closely with AN-speaking people, and northern Chinese ancients were clustered closely with ST-speaking people. We also explored the fine-scale population relationships among Sinitic and TB-speaking people and found Iron Age Hanben people clustered closely with Han Chinese from Guangxi and Taiwan island. However, Yellow River Basin (YRB) farmers clustered separately from other Han Chinese. Clustering patterns among Sinitic and TB people based on Fst matrixes found the differentiated population structure between northern and southern Han Chinese and the genetic differentiation between northern and southern TB people (**Figs. 2D and S3A-B**). We also observed substantial genetic differentiation among Altaic-speaking populations, in which Koreanic and Japonic groups each formed individual clades and separated from other Altaic groups, Mongolic and most Tungusic groups clustered together, and Turkic groups showed a close genetic affinity with some Tungusic groups (**Fig. 2B-C**). However, the fine-scale clustering patterns among AA, AN, HM and TK groups from South China and Southeast Asia indicated extensive gene flow events between these populations. The patterns of phylogenetic relationships and HFS further revealed genetic differences between northern and southern Han Chinese and between northern and southern TB-speaking populations, and gene flow events were identified between geographically close populations, such as between AN and southern Han Chinese and between Altaic and northern Han Chinese (**Fig. 2D**).

We further integrated populations based on their linguistic and ethnic features to explore the genetic affinity among language or ethnicity-based meta-populations via estimating genetic distances and clustering patterns (**Fig. S3C-H**). AN-speaking Saisiyat, Thao, Taroko, Atayal and Tsou

from the island of Taiwan clustered together and separated first from other reference populations (**Fig. S3C**). Other distantly separated branches consisted of populations mainly from TK-speaking people, other geographically close AN (Ede and Giarai) and southern TB (Sila and Lolo) speakers. AN branch and TK-dominant cluster were relatively closer to each other than other Asian reference populations, which provided uniparental genetic evidence for the shared or common origin hypothesis of AN and TK language families. Genetic differentiation between populations from these two branches and others was confirmed via Neighbor-joining (NJ)-based phylogenetic relationship, HFS of major founding lineages, and clustering patterns inferred from the PCA and MDS (**Fig. S3D-F**). Fine-scale genetic differences between Sinitic-related Han Chinese and TB people and genetic distinctions among linguistically different populations were visually visualized here. We should highlight that the phenomena of close relationships among most linguistically distinct populations are common here, suggesting massive population movements and gene flow events occurred in the past. Finally, to assess whether paternal lineages provided some evidence supporting the current consensus of language family classification and explore the genetic relationships between linguistically defined meta-populations, we merged all populations based on their linguistic affinities and conducted population genetic analysis based on the genetic distances and haplogroup frequency distributions (**Fig. S3G-H**). We observed a close clustering relationship between TK and AA based on the genetic distances and NJ-related phylogeny. Our language-defined NJ tree also showed a close relationship between Mongolic, Tungusic and ancient ARB, Turkic and ancient Xinjiang people, Koreanic and Japonic, and AN and ancient Hanben (**Fig. S3H**). We provided robust paternal genetic evidence to support the Chinese people's complex admixture landscape and multifaceted interactions with ancient Eurasians. The simple mergence of geographically distant populations must cause one controversial pattern that may be caused by strategy and sampling biases. We should also pay more attention to the statistically-introduced errors or uncertainties here, which could be gradually overcome in the following WGS-based population genomic studies. We suggested further merging populations by combining the clustering patterns observed in autosomal whole-genome variations and geographical proximity, such as the merged integrated genomic dataset via the administrative divisions.

The Y-chromosomal diversity landscape shaped by complex population migration and admixture events

Patterns of the observed genetic diversity and paternal genetic structure suggested that complex multiple ancient migration and admixture events may have contributed to the formation of the gene pool of Chinese populations. We subsequently systematically tested how many ancestral sources influenced the paternal genetic composition of Chinese populations and explored the geographical distribution of identified lineages and their correlation with ancient pastoralists and farmers' Holocene expansion. The haplogroup information of ancient Eurasian and modern East/Southeast Asian populations was collected to comprehensively characterize the patterns of paternal genetic diversity of ethnolinguistically diverse Chinese populations. The final haplogroup dataset contained 115 ethnically or geographically diverse modern Chinese populations, covering 43 officially recognized or unidentified ethnic groups from all provincial-level administrative divisions except Hong Kong and Macau. Moreover, to explore the geographical origin and distribution patterns of dominant paternal lineages in China, we merged all participants into geographically defined meta-populations and estimated the general geographical distribution patterns (**Figs. 3-4 and S4-7**). There were four dominant paternal macro-haplogroups (C-M130, D1-M174, N-M231 and O-M175) in all included Chinese populations, accounting for about 92% of the Chinese Y-chromosomes; the rest consisted of E-M96, F-M89, G-M201, H-L901, I-M170, J-M304, L-M20, Q-M242, R-M207 and T-M184. Finally, we systematically explored how ancient technological innovation and human migrations influenced the Chinese paternal genetic landscape and identified the flowing four ancient migration and admixture events that enriched Chinese diversity (**Fig. 5**).

The gene flow from ancient pastoralists and barley farmers' paternal lineages in Central/South Asia and West Eurasia to East Asia

Historic and prehistoric tans-Eurasian cultural communication events across the southern Bactrian Marianna archaeological complex (BMAC) oasis farming route, the Inner Asian Mountain Corridor (IAMC) biogeography corridor and the northern Yamanaya/Afanasievo steppe pastoralist-related migration route significantly influenced the autosome-related gene pool of ancient people from Altai mountains and surrounding northwestern and northern East Asia (Zhang, Ning, et al. 2021). Whether the human migrations across the Eurasian Steppe and the

ancient Silk Road introduced Central/South Asian/West Eurasian-derived paternal lineages to present-day Chinese populations remained unresolved. Our work found that haplogroups R, J, G, Q and their major sublineages had the highest frequency in Northwest China and existed in ancient western Eurasians with high frequency (**Figs. 3A-B and S4**). Haplogroup J-M304 most likely evolved in West Asia and was found in significant numbers in present-day West Eurasian populations (Poznik et al. 2016). We observed that most J individuals in China could be subdivided into J2-M172 (especially its sublineage J2a-M410), the distribution center and possible origin of J2a was in the northern Fertile Crescent (**Fig. 3A**) (Grugni et al. 2012), and the current geographical concentration of this lineage largely simulated the agricultural centers (Fuller 2007). Previous findings showed that Northwest Chinese populations, especially Turkic-speaking groups, also carried a strong contingent of J2a (Shou et al. 2010), which was confirmed by our observations. Additionally, ancient individuals belonging to the subclades of J-M304 were identified in the Iron Age Xinjiang populations (Kumar et al. 2022). Chinese individuals carrying G-M201 were also largely restricted to Northwest China. These individuals could be further assigned to G2a (**Fig. 3B**). The spread of G-M201, together with the J2-M172 lineage, was associated with the diffusion of agriculture (**Fig. 3A**) (Semino et al. 2000). Estimates of optimized hot spot analysis (OHSA) confirmed that the diffusion centers of J2a and G2a in China were correlated to Xinjiang and Gansu-Qinghai regions (**Fig. S4A**). Generally, the J/G-derived lineages were likely introduced into China during the eastward migration of Central Asian-related ancestral populations, which may result from the gene flow events mediated by the ancient Silk Road (Zhabagin et al. 2022; He, Yao, et al. 2023).

Haplogroup R-M207 was distributed in ancient western Eurasian and also in modern people in North China at a relatively high frequency, especially in Northwest China (**Fig. 3A-B**), which has been reported to have emerged and diversified in Central Asia and is now common throughout Central/South Asia and Europe (Kayser et al. 2003; Hallast et al. 2021). Only an approximately 24,000-year-old individual (MA1) from the Mal'ta site near Lake Baikal in Siberia belonged to basal haplogroup R (Raghavan et al. 2014). We found that most R individuals could be subdivided into R1-M173, and the haplogroup frequency of its subclade R1a-L146 in China was much higher than that of R1b-M343. In addition, a majority of R1a individuals could be subdivided into R1a1a sublineages, and an extended Y-chromosome investigation revealed R1a1a-M17 as one of the paternal lineages that entered East Asia via the northern route (Zhong et al. 2011). R1b subclades are frequently found in Western Europe (**Fig. 3A**). The estimation of the spatiotemporal distributions of R-M207 subclades supported a West Asian origin of R1b and a subclade carrying the M269 mutation (R1b1a1b-M269) quickly spread to Europe and diffused across Western Europe (**Fig. 3A**) (Myres et al. 2011; Olalde et al. 2018). Haplogroup R2-M479 was found in East Asia at an extremely low frequency, which has been geographically concentrated in Central/South Asia and recently spread from South Asia to North China via the northern route (Zhong et al. 2011; Di Cristofaro et al. 2013). Ancient DNA studies showed that several Mongolia_EIA_Sagly_4 and Mongolia_LBA_MongunTaiga_3 samples that carried more than 40% West Eurasian-related ancestry belonged to R1a1a sublineages (Wang, Yeh, et al. 2021) and several Bronze and Iron Age Xinjiang individuals belonged to R1b sublineages (**Fig. 3A**) (Zhang, Ning, et al. 2021; Kumar et al. 2022). Our findings were consistent with the high frequency of these paternal lineages identified in ancient pastoralists of Yamanaya and Afanasievo populations in the western Eurasian Steppe (Narasimhan et al. 2019; Kumar et al. 2022). The influence of Central/South Asian/West Eurasian-related ancestors could also be reflected by the sporadic occurrence of minor haplogroups, such as E-M96, F-M89, H-L901, I-M170, L-M20 and T-M184, which generally revealed relatively extensive and recent gene flow from these ancestral populations to China.

To further confirm the pastoralist-related population migrations were conferring the reshaped patterns of western Eurasian-related lineages in our studied populations, we estimated the correlation between haplogroup frequency and geographical (longitude and latitude) and genetic features (PC1-2, haplogroup frequency, Fst matrix and autosome-based admixture proportion under the best-fitted models). The frequency of R-related lineages was correlated with latitude and strong genetic affinity with northwestern modern and ancient Chinese populations (**Fig. 4A-B**). R lineages and their sublineages were also observed with a strong correlation (**Fig. 4C**). To illuminate the direct genetic contributions from ancient sources to modern people, we constructed one best-fitted six-source admixture model to explore the ancestral components and proportions of eastern Eurasians and found ancestral proportion decreased gradually from their archeologically attested origin centers (**Fig. 5A-G**). If ancient population movement events directly dispersed the patterns of lineage frequency of ethnolinguistically diverse modern populations, we expected to observe a strong positive

correlation between autosome-based admixture proportion of the putative ancestral source and the frequency of the founding lineages. Interestingly, we found a strong correlation between Afanasievo-related ancestry and multiple C, J, N and R sublineages (**Fig. 5H and 5N**). Our results from the HFS of modern and ancient populations, phylogenetic origin inference and multiple factor correlations suggested that western barley and pastoralist people may have promoted the formation of the aforementioned founding lineages.

The dominant Siberian hunter-gatherers' paternal lineages are widely distributed in China

Ancient DNA studies identified an ancestral component that represented the lineage related to Neolithic hunter-gatherers who once lived in the Russian Far East, the Baikal region and Mongolian Plateau, which was referred to as Ancient Northeast Asian (ANA) ancestry (**Fig. 5A and 5C**) (Jeong et al. 2020; Mao et al. 2021). We found that the ANA ancestry made varying contributions to the surrounding spatiotemporally distinct ancient populations, who harbored high proportions of haplogroups Q, R, C and N (Jeong et al. 2020; Wang, Yeh, et al. 2021; He, Yao, et al. 2023). These lineages in the gene pools of modern Mongolic, Tungusic, Turkic and other groups had a significant positive correlation with the proportion of ANA-related ancestry (p values less than 0.05, **Fig. 5N**). Haplogroup Q-M242 was scattered in Chinese populations at extremely low frequencies and its subclades showed different distribution patterns in North and South China (**Fig. 3C**). This lineage might have originated in Central Asia and southern Siberia 15,000–25,000 years ago. Then, its subclades continued to spread from these regions over the past 10,000 years (Huang, Pamjav, et al. 2018). A major subclade, Q1a1a-M120, was unique to East Asians and found in Han Chinese at a relatively high frequency, which proved to be a crucial lineage of Han Chinese and underwent an *in-situ* expansion in Northwest China between about 5,000 and 3,000 years ago (Sun et al. 2019). Ancient genomes revealed that most of the Ulaanzuukh_SlabGrave individuals and a Mongolia_LBA_CenterWest_4 individual who carried a minimum proportion of West Eurasian-related ancestry (< 20%) belonged to Q1a1a or its sublineages (**Fig. 3A**) (Wang, Yeh, et al. 2021). Venn-based shared ancestry-correlated lineages also revealed that Q and R lineages were the shared lineages among Yamnaya and ANA-associated lineages (**Fig. 5O**). Besides, one middle Neolithic Yangshao individual and some ~3,000-year-old Hengbei people from Shanxi belonged to Q1a1a-M120 (Zhao et al. 2014; Ning, Li, et al. 2020), suggesting that ancient individuals carrying Q1a1a contributed to Han Chinese at least 6,000 years ago. Haplogroup Q1b-M346 was rare in China and distributed most intensively in the crossroad between Siberia and North China (**Fig. S4B**), which was reported to have a wide distribution across Asia (Huang, Pamjav, et al. 2018). Some Bronze and early Iron Age individuals from Mongolia were genotyped Q1b1 sublineages (Wang, Yeh, et al. 2021), and some Bronze and Iron Age Xinjiang individuals belonged to Q1b or its subclades (**Fig. 3A**) (Zhang, Ning, et al. 2021; Kumar et al. 2022).

One of China's most prevalent paternal lineages is haplogroup N-M231, especially its subclade N1-CTS3750. Genetic evidence suggested that N-M231, which presumably originated in South China, spread northward into North China and Siberia during the late Pleistocene-Holocene as the climate warmed and then migrated westward on a counter-clock path from southern Siberia/Inner Asia (Rootsi et al. 2007; Shou et al. 2010; Zhong et al. 2011). Cui et al. conducted Y-chromosome analyses on ancient individuals dating to 6500 to 2700 BP from the West Liao River (WLR) Basin. They found that N-M231 was the leading paternal lineage in Northeast China in the Neolithic Age, and its haplogroup frequency declined gradually over time (Cui et al. 2013). We found that N1a-F1206 showed a relatively high frequency in North China, while N1b-F2930 showed a relatively high frequency in South China, especially in low-altitude Southwest China (**Fig. S4B**). Previous findings provided evidence for the differential distribution patterns of N1a and N1b. That is, the differentiation of the two may have occurred in North China, and then N1a-F1206 migrated northward to areas outside East Asia, while N1b-F2930 migrated southward to South China and became one of the major paternal lineages of TB groups, especially Yi people (Rootsi et al. 2007; Shi et al. 2013; Ilumae et al. 2016; Wang, Song, et al. 2021). N1a1-M46/Tat was a dominating subclade of N1a that probably originated in Northeast Asia (Rootsi et al. 2007; Hu et al. 2015). A sample from the Houtaomuga site in Jilin dating to 7430–7320 years ago belonged to the subclade of N1a1a1a1a-M2117, which was genetically related to early Neolithic ARB individuals, such as DevilsCave_N and Boisman_MN (Ning, Fernandes, et al. 2020). Wang et al. found that several Bronze and Iron Age individuals from the ARB and Mongolia were assigned to N1a or its sublineages, such as a Yankovsky_IA sample (I1202, N1a) and a Mongolia_EIA_SlabGrave_1 sample (I6365, N1a1a1a1a) (Wang, Yeh, et al. 2021). N1a2-F1008/L666 is a dominant paternal type of Uralic populations (Hu et al. 2015), and an ancient

DNA study showed that three early Neolithic Shamanka individuals from the Cis-Baikal area were found to belong to N1a2-L666 (de Barros Damgaard et al. 2018). We found that North China and the southwestern part of Northeast China might be the early diffusion center of N1a2 (**Fig. S4B**), which was generally consistent with previous observations (Yu et al. 2023). We observed that N1b-F2930 was mainly distributed in TB-speaking populations in Southwest China and found infrequently in other Chinese populations. However, none of the reported ancient East Asians belonged to haplogroup N1b, and the fine-scale phylogenetic structure of N1b-F2930 needs to be further explored.

Haplogroup C-M130 is a major paternal lineage in East Asia, which might be carried by one of the waves of the earliest settlers, and its diffusion in East Asia was speculated to have begun about 40 thousand years ago (kya) (Ke et al. 2001; Zhong et al. 2010; Wang and Li 2013). C2-M217 was the most widespread subclade and was found with a high frequency in North China (**Figs. 3A, 3C and S4B**). The oldest individual carrying C2-M217 identified so far was AR19K (19,587-19,175 cal BP) in the ARB (Mao et al. 2021). We observed distinct distribution patterns of C2a-L1373 and C2b-F1067, with the highest frequency of C2a-L1373 (sometimes referred to as the "northern branch" of C2-M217) occurring in the Inner Mongolia Autonomous Region and the highest frequency of C2b-F1067 (sometimes referred to as the "southern branch" of C2-M217) in Central, North and Northeast China. We found that most C2a individuals could be further assigned to C2a1a, and its sublineages were often found in Altaic-speaking populations. Specifically, C2a1a1b1-F1756, C2a1a2a-M86 and C2a1a3a-F3796 were the major subclades of C2a1a individuals in China (**Fig. S4B**). The phylogeny reconstruction of C2a1a1b1-F1756 suggested that its two major subclades might be related to the early expansions of Mongolic/Tungusic-related ancestors (Wei et al. 2017). Several studies have shown that C2a1a2a sublineages were widespread in Altaic groups in East Asia and North Asia (Chen et al. 2011; Liu, Ma, et al. 2021; Wang, He, et al. 2021), and some ancestral Tungusic groups in the ARB migrated to the Mongolian Plateau and contributed the genetic components of C2a1a2a to present-day Mongolic/Turkic speakers (Liu, Ma, et al. 2021). C2a1a3a-F3796, also known as the C2*-Star Cluster (C2*-ST), is one of the founding paternal lineages of Mongolic-speaking populations (Zerjal et al. 2003; Wei et al. 2018; Wang, He, et al. 2021). The highly revised phylogenetic tree of C2*-ST and the estimated time of TMRCA of this paternal lineage and its sublineages suggested that the dispersal patterns of C2*-ST correlated with the expansion of Mongolic-speaking populations, whose origins could be traced back to either ordinary Mongolian tribes or an ancient Niru'un Mongol clan (Wei et al. 2018). Wang et al. found that several ancient individuals from the ARB and Mongolia belonged to C2a1a or its sublineages, such as Mongolia_North_N and Boisman_MN (**Fig. 3A**) (Wang, Yeh, et al. 2021), indicating that ANA-related populations contributed significantly to present-day Mongolic, Tungusic and some Turkic groups. We also identified some individuals belonging to C2a sublineages in central and southern Chinese populations, indicating that Proto-Mongolian populations from North Asia migrated southward and spread the C2a sublineages, which was mainly driven by the expansion of the Mongol Empire (Wei et al. 2017; Zhang, Wu, et al. 2018; Li et al. 2020).

The phylogenetic analysis of C2b-F1067 suggested that ancient populations carrying various C2b sublineages contributed significantly to the gene pool of modern Eastern Eurasians (Wu et al. 2020). Our observations showed that the Inner Mongolian Plateau and Northeast China might be the initial dispersal center of C2b (**Fig. S4B**), consistent with a previous finding that North China might be the diffusion center of C2b before 11 kya (Wu et al. 2020). We revealed different geographical distribution patterns of C2b sublineages (**Fig. S4B**). For example, C2b1a1-CTS2657 was distributed in North and Northeast China at relatively high frequencies; C2b1a2-F3880 was prevalent in Northeast, North (except the Inner Mongolia Autonomous Region and Shanxi) and East (mainly Shandong, Jiangsu and Shanghai) China; while C2b1b-F845 has distributed in Central and Southwest (mainly Guizhou) China as well as Southeast Asia at the highest frequency. Significantly, our findings were partially inconsistent with a previous study (Wu et al. 2020), which may be caused by the bias of sampling and reference populations. Our study and several previous studies confirmed the southern origin of C2b1b-F845, and an ancient DNA study further suggested that the origin of this paternal lineage was associated with southern farming populations, who contributed the genetic components of C2b1b to ancient nomadic groups on the Mongolian Plateau (Li et al. 2020; Wu et al. 2020). Statistically significant negative correlations between western Eurasian and Siberian-related lineages suggested that their high frequency in northern East Asians contributed to their genetic differentiation from southern East Asians (**Fig. 4B**). Generally, genetic analyses based on the modern and ancient East Asians suggested the solid genetic connection between Neolithic Mongolian hunter-gatherers and

modern East Asians.

Early Asian paternal lineages left traces mainly in the Tibetan Plateau and surrounding regions

The ancient genetic connection between Andamanese, Jomon-related indigenous Japanese, and highland Tibetans was evidenced via the shared Palaeolithic autosomal ancestry components and uniparental D lineage (**Fig. 3A and 3D**). We explored the phylogeographical origins of D subclades and found that D1-M174, as one of the four major Y-chromosome haplogroups in East Asians, did occur at high frequency in our YHC (**Fig. S5**), and this lineage had been intensively studied and showed clear southern origin (Shi et al. 2008; Zhong et al. 2011; Qi et al. 2013). Haplogroup D1a exhibited high proportions and clustering centers in the Tibetan Plateau (TP) and the Japanese archipelago (**Fig. S5**). Significantly, most D1a individuals (> 95%) could be subdivided into D1a1a-M15 or D1a1b-P99. We observed that D1a1a sublineages were found frequently among TB-speaking populations in Southwest China and at low frequency in other Chinese populations, whereas D1a1b sublineages occurred at the highest frequency on the TP. Previous studies showed that D1a1a-M15 diverged from D1-M174 during the migration of D1-M174 to East Asia, and then D1a1a-M15 migrated northward through western Sichuan to Gansu-Qinghai region and might have entered the Himalayan area along the Tibetan-Yi corridor; D1a1b-P99, primarily its subclade D1a1b1-P47, diverged from D1-M174 on the TP, which resulted in D1a1a sublineages being widely distributed in China and D1a1b sublineages being specific to Tibetan populations (Qi et al. 2013; Wang et al. 2014). Sublineages of D1a were uniquely maximized in Tibetan populations, which was confirmed by the genetic contribution from North China millet farmers to highlanders via the revised Y-chromosome phylogeny, correlation with O lineages and Lubrak-related TP ancestry (**Fig. 4C and 5D**). The estimated gene flow events, Lubrak-related D sublineages and the close genetic connections between North China and TP highlanders contributed to Chinese people's genetic diversity patterns. Interestingly, the frequency of four lineages (O2a2b1, O2a2b1a, O2a2b1a1 and O2a2b1a1a) also strongly correlated with the Lubrak-related ancestry further confirmed the Neolithic expansion from YRB promoted the peopling of the highland TP (**Fig. 5J**).

Ancient northern East Asian millet farmer-related farming-language-people southward dispersal

Archeological and historical documents suggest that YRB was the cradle of ancient Chinese civilization. Ancient DNA of millet farmers from Houli, Yangshou and Longshan cultures showed that all ST people originated from North China (Wang, Yeh, et al. 2021). Wang et al. investigated the genetic profiles of spatiotemporally different Guangxi populations and identified southward genetic influence related to Shandong ancients (Wang, Wang, et al. 2021). Yang et al. identified persistent southward gene flow from Shandong to the coastal southeastern East Asia (Yang et al. 2020). We also identified Haojiatai-related ancestry dominant in Chinese populations, strongly correlated with the O, Q, C and N lineages. Haplogroup O-M175 is primarily found in East and Southeast Asians and has two main subclades: O1-F265 and O2-M122. The origin, diversification and expansion of these two lineages and their sublineages were suggested to be associated with the spread of millet and rice farmers from the domestication agriculture centers of the YRB and Yangtze River Basin (YZRB) (**Fig. 5K-M**). However, the extent to which ancient northern East Asians reshaped the paternal genetic diversity of modern southern East Asians remained to be comprehensively assessed. Thus, we evaluated the geographical distribution of O-related lineages and explored the correlation with the estimated millet farmer ancestry (**Fig. 5K**). Ancient O2 lineages were widely distributed in northern China and the TP (**Fig. 3A**). Our findings showed that haplogroup O2-M122, especially its subclade O2a-M324, is one of the major paternal lineages in East Asian populations and is also highly prevalent in Southeast Asian populations (**Fig. 3D and S6**) and a strong correlation within different subclades (**Fig.4C**). Y-chromosome evidence based on 15 Y-SNPs demonstrated a southern origin of O2-M122 and suggested that its northward migration in East Asia occurred ~25–30 kya (Shi et al. 2005). We observed that O2a-M324 was widely distributed in China's coastal and surrounding areas with extremely high frequency (**Fig. S6B**), indicating that the ancestors carrying O2a-M324 were likely to migrate along the coast and then spread to other East Asian areas and Southeast Asia. Yan et al. sequenced 3.9 Mbp NRY region of 78 East Asians and identified three star-like Neolithic expansions [Oα (O2a2b1a1-M117), Oβ (O2a2b1a2a1a-F46) and Oγ (O2a1b1a1a1a-F11)] in haplogroup O2a-M324, which suggested that the male-mediated expansion in China mainly occurred during the Neolithic Age (Yan et al. 2014). Ancient DNA evidence revealed that a middle Neolithic individual from the WLR belonging to Hongshan culture was assigned to haplogroup O2a-M324 (Ning, Li, et al. 2020). Yu and Li

reviewed the origin of Chinese ethnic groups, language families and civilizations from the perspective of Y-chromosome and revealed that O2a-M324 originated in northeastern China and was associated with the development of Hongshan culture. We did observe that the highest frequency of haplogroup O2a-M324 occurred in northeastern China's Heilongjiang (**Fig. S6B**). However, its highest haplogroup frequency was also observed in some eastern coastal provinces, such as Shandong, Shanghai, Fujian and Guangdong, which may be caused by the bias in included ethnically diverse populations and sample sizes. Additionally, OHSA results indicated that the middle and lower YRB could be the early diffusion center of O2a-M324.

We observed distinct distribution patterns of O2a1-L127.1 and O2a2-JST021354/P201 (**Fig. S6**). O2a1 occurred at the highest frequency in East China, and its haplogroup frequencies decreased in the surrounding areas (**Fig. S6C**). At the same time, O2a2a-M188 was distributed in Southeast Asia at relatively high frequencies and then spread from south to north in East Asia with a decrease in the haplogroup frequency. O2a2b-P164 was prevalent in China and occurred at the highest frequency in the TP. We found that the overwhelming majority of O2a1 individuals could be further subdivided into O2a1b-JST002611, and it was widely distributed in different Chinese ethnic groups, especially in Han populations (**Fig. S6D**), which was consistent with previous genetic findings (Wang et al. 2013; Yao et al. 2017). The low-frequency distribution of O2a1b and its sublineages in TB-speaking populations suggested that this lineage did not play a significant role in forming TB speakers. There are two main sublineages in O2a1b defined by F11 (O2a1b1a1a1a) and F238 (O2a1b1a2a), respectively (Wang et al. 2013; Yao et al. 2017). We observed that O2a1b1a1a1a-F11 was more frequently distributed in China than O2a1b1a2a-F238, especially in geographically diverse Han populations. Haplogroup O2a1b1a1a1a was distributed in East China at higher frequencies, which was consistent with previous findings that O2a1b1a1a1a-F11 experienced a rapid expansion probably in eastern Han Chinese about five kya (Wang et al. 2013). Wang et al. found that the F238 mutation was likely to occur in Proto-Han-Chinese about seven kya, but it was difficult to clarify the origin of this mutation based on the Y-STR haplotype diversities of O2a1b1a2a-F238 (Wang et al. 2013), which was confirmed by the observation in this study that haplogroup O2a1b1a2a was mainly restricted to Northeast, North and East China (**Fig. S6E**). We found that the initial diffusion center of O2a1b sublineages was likely to be the middle and lower YRB (**Fig. S6D**). Zhang et al. reported genome-wide data from 12 ancient samples dating to 6500 to 2500 years ago in China and identified an O2a1b1a1a1a-F11 sample from the Banpo site, indicating that the Yangshao people may have induced the generation of haplogroup O2a1b1a1a1a (Zhang, Lei, et al. 2018). Ning et al. found that a late Neolithic individual from the Erdaojingzi site belonging to the Lower Xiajiadian culture in the WLR (WLR_LN) could be assigned to haplogroup O2a1b-JST002611 and WLR_LN derived their ancestry mainly from YRB-related ancestral populations (88 or 74%) (Ning, Li, et al. 2020).

Most O2a2a subclades were distributed with high frequencies in South China and Southeast Asia (**Fig. S6F-G**). Previous studies suggested that O2a2a1a2-M7 was one of the founding lineages of HM-speaking populations (Xia et al. 2019; Kutanan et al. 2020; Liu, Xie, et al. 2021) and found at a high frequency in the Daxi people in the middle reaches of the YZRB (Li et al. 2007), demonstrating that the Daxi people might be the ancestors of present-day HM speakers. We found that O2a2b sublineages were widely distributed in China. O2a2b1-M134 was one of the main subclades of O2a2b, which was found frequently among ST-speaking populations, and the highest frequency occurred in Tibetan populations (**Fig. S6H**). Two O2a2b1 subclades (O2a2b1a1-M117 and O2a2b1a2a1a-F46) showed star-like expansions (Yan et al. 2014). The highest frequency of O2a2b1a1 was observed in the TP and Southeast China, but the OHSA results showed that the early diffusion center of O2a2b1a1 might be the TP (**Fig. S6H**). We found that O2a2b1a2, the upstream haplogroup of O2a2b1a2a1a, was distributed in Northeast, North and East China at higher frequencies than other regions (**Fig. S6H**). Ancient DNA studies revealed that O2a2b1 sublineages were identified in ancient individuals from the Shimao site belonging to the Longshan culture (Shimao_LN), from the Jinchankou, Lajia and Mogou sites belonging to the Qijia culture (Upper_YRB_LN) and from the Dacaozi site (Upper_YRB_IA) (Li et al. 2017; Ning, Li, et al. 2020). Wang et al. demonstrated that O2-M117-F5 (called Oα-F5) was one of the founding paternal lineages of modern TB groups and the Yangshao people in the middle YRB carrying Oα-F5 migrated southwestward to the TP at about six kya and mixed with the local D-M174 populations to form the first layer of the genetic profiles of modern TB-speaking populations (Wang et al. 2018). Furthermore, several genome-wide SNP studies revealed that Core Tibetans (Tibetans from the TP) had more than 70% ancestry related to Upper_YRB_LN or more than 85% ancestry related to YRB_MN (Wanggou_MN and Xiaowu_MN) belonging to the Yangshao culture (He, Wang, Zou, et al. 2021; Wang, Yeh, et al. 2021). Therefore, the development of millet

agriculture, migration of millet farmers and admixture with geographically diverse indigenous populations resulted in the current distribution patterns of diverse O2a-M324 sublineages, with the leading subclades of O2a1b-JST002611 and O2a2b1-M134.

Ancient southern East Asian rice farmer-related founding lineages from the YZRB left a massive genetic legacy in China and Southeast Asia

South China was another agricultural origin center of rice domestication and was proposed as the hometown of HM, TK, AA and AN people. The best-fitted ADMIXTURE models revealed inland Hmong- and coastal Hanben-related ancestral components widely distributed in southern Chinese populations and correlated with most O1 subclades (**Fig. 5F-G and 5L-O**). Recent ancient DNA also illuminated that ancient YZRB rice farmers genetically reshaped the patterns of ancient YRB millet farmers and Southeast Asian modern and ancient populations (Yang et al. 2020; Wang, Wang, et al. 2021). We here explored the phylogeographical features of O1 sublineages to illuminate the impact of paleo-genomically attested gene flow events on the paternal genetic diversity of Chinese and Southeast Asians. We found that O1-F265 was found at high frequencies in South China, Southeast Asia and the Japanese archipelago, and its subclade O1a-M119 was frequently found in southeastern Chinese populations, while O1b-M268 was frequently distributed in southwestern Chinese and Southeast Asian populations (**Fig. S7**). O1a sublineages were mainly distributed in AN-, TK- and Sinitic-speaking populations in Southeast Asia and South China, indicating that AN and TK speakers shared a common patrilineal ancestor and extensive gene flow existed between Han Chinese and AN/TK-related ancestral populations. Their genetic interaction and admixture signatures were also identified via the autosome-based admixture models (Chen et al. 2022; Wang, He, et al. 2022; Liu et al. 2023). We observed that O1a1a1 and its sublineages (except O1a1a1b) were found at high frequencies in Southeast China, and their early dispersal centers might be the middle and lower reaches of the YZRB and the southeast coast (**Fig. S7B**). Previous research showed that the paternal lineages observed in the Liangzhu populations from the Yangtze River Delta were mostly O1a-M119 and in the Taiwan Hanben individuals were mostly O1a1 sublineages (Li et al. 2007; Wang, Yeh, et al. 2021), and rice farmers carrying O1a-M119 in the YZRB were likely to be the direct ancestors of modern TK and AN speakers, they migrated southward along the southeast coast of China to Southeast/Southwest China and mainland Southeast Asia (Liu et al. 2020; Wang, Yeh, et al. 2021; Wang, Wang, et al. 2021; Wang, He, et al. 2022). We observed that O1a1a1b was found at the highest frequency in Hainan, especially in Li people, and its haplogroup frequency showed a general south-to-north decline, suggesting that Li-related ancestors made significant contributions to other southern Chinese populations (**Fig. S7B**). Our observations showed that O1a1a2 and its sublineages were found at high frequencies in Southwest China and Vietnam, and Southwest China was likely to be the initial diffusion center of these paternal lineages (**Fig. S7C**). However, the high-resolution phylogeny of O1a-M119 revealed that O1a1a2-F4084+/K644- were found at higher frequencies in the Yangtze River Delta, while its southwestern sublineage O1a1a2a1a-K644, a founding paternal lineage of TK-speaking populations, showed an apparent south-north dispersal trend starting from Hainan Island (Sun et al. 2021). Substantial differences in the coverage of target populations and genotyping methods may lead to partial inconsistencies in the results of different studies. Generally, the migrations of rice farmer-related ancestral populations largely contributed to the observed distribution patterns of O1a sublineages.

We observed that O1b-M268 was distributed in Southwest China, Southeast Asia and the Japanese archipelago at high frequencies, which could be subdivided into three major subclades (O1b1a1-PK4, O1b1a2-Page59 and O1b2-P49) that showed significantly different distribution patterns (**Fig. S7D**). O1b1a1 and its sublineages were mainly distributed in Southwest China and Southeast Asia, and a significant sublineage O1b1a1a-M95 mainly existed in AA groups in these regions and was also an important paternal lineage of TK-speaking populations (Zhang et al. 2014; Kutanan et al. 2019; Song et al. 2019; Macholdt et al. 2020). However, the geographical origin and migratory routes of O1b1a1a remain controversial. Ancient DNA evidence revealed that the Wucheng people in Jiangsu along the YZRB and two ancient individuals from the Hengbei site in Shanxi approximately 3,000 years ago carried the O1b1a1a-M95 lineage (Li et al. 2007; Zhao et al. 2014). We observed that O1b1a2 and its sublineages were relatively rare in East Asia and mainly distributed in East China, the southwestern part of Northeast China and Vietnam, especially in Han Chinese (**Fig. S7E**), consistent with previous findings (Yan et al. 2011). Ancient genomes from North China revealed that a middle Neolithic individual from the Wanggou site belonging to the Yangshao culture was assigned to O1b1a2-Page59 (Ning, Li, et al. 2020). Haplogroup O1b2-P49 was found at the highest frequency in Japan, followed by Northeast China, but the detailed

phylogenetic structure of this lineage remains to be further reconstructed (**Fig. S7F**). Our identified patterns of genetic diversity from O1 lineages suggested that ancient rice farmers from South China significantly influenced the gene pool of populations from South China and Southeast Asia. In general, complex population movements and admixture events contributed to the formation of modern and ancient East Asians. To illuminate the origins of different Chinese-dominant subclades and the demographic processes of ethnically/geographically diverse modern Chinese populations, we should design a systematic sampling strategy, conduct whole Y-chromosome sequencing and fully retrieve spatiotemporally distinct ancient individuals for more comprehensive analyses.

Conclusion

Genetic evidence from autosome-based ancient DNA has revolutionized our understanding of human population genetic history; however, ancient genetic legacy inferred from the ancient Y-chromosome was limited, as the lower copy number than mitochondrial genomes. We launched the YanHuang cohort to capture the Y-chromosome diversity of ethno-linguistically diverse Chinese populations via genotyping 919 individuals from 39 ethnic minority groups using our recently developed high-resolution YHSeqY3000 panel and merged with the Y-chromosome genomic database of 14,611 people, including 1753 ancient people, to explore the formation process of ancient Chinese Y-chromosome genetic diversity landscape. Ancient DNA data collected from ancient Eurasian populations and modern integrated data included 115 ethnolinguistically or geographically distinct modern Chinese populations belonging to 47 officially recognized or unidentified ethnic groups and covered all provincial-level administrative divisions except Hong Kong and Macau. Our results identified multiple founding lineages related to ancient western herder Euraisan, Siberian Fisher-Hunter-Gatherer and millet and rice farmers from Yellow and Yangtze River Basians contributed to the patterns of geography-related population paternal genetic stratification. We illuminated the strong correlation between the frequency of subsistence-model-related founding lineages and the autosome-based admixture proportion of putative ancestral sources, latitude and differentiated north-to-south genetic matrix, suggesting ancient population movements and extensive admixture between incomers and indigenous populations was the major mechanism for the formation of the evolutionary spectrum of the Y-chromosome landscape. We emphasized the importance of combining high-deep whole-genome sequencing data of modern and spatiotemporally different populations to validate and further characterize the paternal evolutionary history of East Asians.

Materials and Methods

Study participants

To fully characterize the panorama of Y-chromosomal diversity in China, we collected saliva samples from 919 participants from 39 ethnolinguistic groups belonging to Sinitic (Hui), TB (Bai, Derung, Lahu, Lisu, Lhoba, Nakhi, Pumi, Qiang, Tibetan, Tujia and Yi), TK (Bouyei, Dai, Dong, Gelao, Li, Maonan, Shui, Zhuang and Mulao), HM (Miao, She and Yao), AN (Gaoshan), AA (Jing and Wa), Mongolic (Daur, Dongxiang and Mongolian), Tungusic (Hezhen, Manchu and Xibe), Turkic (Kazakh, Kyrgyz, Uzbek and Uyghur), Koreanic (Korean) and IE (Russian) language families (**Table S1**). The sampled individuals were descendants of self-identified members of the given ethnic groups and their grandparents lived in the sampling districts for at least three generations. This study was approved by the Medical Ethics Committee of West China Hospital of Sichuan University (2023-306) and conducted in accordance with the Helsinki Declaration of 2013 (Jama 2013). In addition, we obtained informed consent from each participant.

DNA extraction, sequencing and genotyping

We extracted genomic DNA using the QIAamp DNA Mini Kit (QIAGEN, Germany). DNA concentrations were quantified using the Qubit dsDNA HS Assay Kit based on the standard protocol on a Qubit 3.0 fluorometer (Thermo Fisher Scientific). The Y-specific target sequences with 50X coverage were generated on the Illumina platform (Illumina, San Diego, CA, USA) using the custom-designed YHSeqY3000 panel. The raw sequencing reads were mapped to the human reference genome GRCh37 using BWA v.0.7.13 (Li and Durbin 2009).

Haplogroup classification and phylogeny construction

We first conducted NRY haplogroup classification using in-house scripts. The NRY haplogroups were also classified using HaploGrouper (Jagadeesan et al. 2021) and Y-LineageTracker (Chen et al. 2021) based on the Y-DNA Haplogroup Tree 2019–2020 (https://isogg.org/tree/index.html), respectively. For the complete dataset of 919 samples, a maximum-likelihood phylogenetic tree was constructed via MEGA X (Kumar et al. 2018), which was then visualized using iTOL (Letunic

and Bork 2021). The haplotype and haplogroup diversity were estimated using the following formula: HD/H=$\frac{N}{N-1}(1-\sum Pi^2)$, where N denotes the total number of observed haplotypes or haplogroups and Pi denotes the frequency of the i-th haplotype or haplogroup. We only extracted populations with a sample size greater than 30 for the estimation of HD and H values.

Haplogroup frequency spectrum estimation and clustering analysis

Dataset composition

We incorporated previously published haplogroup information of 11,979 East Asian individuals from 79 populations retrieved from previous fragmentation studies (Trejaut et al. 2014; Lang et al. 2019; Song et al. 2019; Xie et al. 2019; Wang, Song, et al. 2021; Wang, He, et al. 2021; Song et al. 2022; Wang, Song, et al. 2022), the 1KGP and the Human Genome Diversity Project (HGDP) (Poznik et al. 2016; Bergstrom et al. 2020), 879 individuals from 28 Southeast Asian populations (Kutanan et al. 2019; Kutanan et al. 2020; Macholdt et al. 2020), 131 ancient East Asians from Xinjiang (Xinjiang_China), ARB (AmurRiver_China), YRB (YRB_China) and South China (Hanben_IA) (Jeong et al. 2020; Wang, Yeh, et al. 2021; Zhang, Ning, et al. 2021; Kumar et al. 2022) and 1622 ancient western Eurasians from Allen Ancient DNA Resource into this study to estimate the HFS and depict the population structure of ethnolinguistically diverse Chinese ethnic groups. A total of 13,777 present-day individuals were collected from 12 linguistically different groups, covering 22 provinces, five autonomous regions and four municipalities in China as well as Thailand and Vietnam, including 135 AA-, 693 AN-, 285 HM-, 75 Japonic-, 35 Koreanic-, 994 Mongolic-, 863 TK-, 1338 TB-, 260 Tungusic-, 291 Turkic-, 1 IE-speaking individuals (excluded from population genetic analysis), 805 Sinitic-speaking Hui, 3248 northern Han Chinese and 4754 southern Han Chinese (**Tables S1 and S3**). The retrieved haplogroups were manually revised based on the variant information and the Y-DNA Haplogroup Tree 2019–2020. In order to estimate the spatial distributions of different paternal lineages more conveniently, we integrated haplogroup data to generate meta-populations based on the geographical region, ethnicity and language family. We estimated haplogroup frequency based on different levels of terminal haplogroups. We conducted population genetic analysis based on individual populations with a sample size over ten or meta-populations with a sample size larger than 30. Haplogroup frequency of all upstream lineages was used to explore corresponding deep population genetic history.

Population structure inference

We calculated pairwise Fst genetic distances based on the HFS using Y-LineageTracker (Chen et al. 2021). The NJ phylogenetic tree was constructed using MEGA (Kumar et al. 2016) and modified using the iTOL online tool. MDS was conducted based on the genetic distance matrix using the cmdscale R tool (https://itol.embl.de/itol.cgi), and PCA was conducted based on the HFS using Y-LineageTracker (Chen et al. 2021).

Spatial statistics correlated the phylogeographical origin of founding lineages

The haplogroup frequency of one province-defined population at different levels of terminal haplogroup was calculated using Y-LineageTracker with level parameters ranging from 0 to 6. Geographically diverse Chinese populations were merged based on the administrative distinctions of provinces, and populations from the island and mainland of Southeast Asia were integrated based on the country. Using ArcMap, we investigated the geographical distribution patterns of dominant haplogroups in Chinese populations by OHSA (Getis-Ord General G) and spatial autocorrelation analysis (Moran's I). The clusters (hot and cold spots) revealed by OHSA roughly mirrored the possible geographical origin or diffusion center of a specific haplogroup and its mirror regions showed the general distribution pattern of that haplogroup.

Phylogenetic relationship reconstruction

Fasata data was used to reconstruct the phylogenetic tree using using Y-LineageTracker (Chen et al. 2021). Network relationship of shared haplotype was explored via the popart (Leigh et al. 2015) software.

Autosome-based ADMIXTURE estimation

We collected 445 ancient individuals from 88 Eurasian populations and 1325 geographically different modern individuals from 62 populations from our merged 10K_CPGDP database to form the autosome-based dataset, including ancient Chinese populations from inland and coastal southern and northern East Asia, hunter-gatherers from Mongolia and western Yamnaya-related pastoralists. We modeled the admixture proportions of geographically different populations via best-fitted ADMIXTURE. We pruned the autosome-based dataset using PLINK (Chang et al. 2015) with the parameters of "--indep-pairwise 200 25 0.4" and "--allow-no-sex" and then ran ADMIXTURE with the predefined ancestral sources ranging from 2 to 15 (Alexander et al. 2009). We used cross-validation error values to identify the best-fitted admixture models and used the

admixture proportion of modern populations to correlate autosome-based ADMIXTURE and haplogroup frequency.

Correlation between lineage frequency and ADMIXTURE-based ancestry proportion

We first calculated the haplogroup frequency of geographically defined meta-populations. Chinese populations were grouped based on the provincial administrative region. We cut all tested lineages at the nine levels and identified 139 common lineages with a frequency larger than 0.05 in at least one population, 177 low-frequency lineages and 165 rare lineages. We then explored the Person correlation between haplogroup frequency and longitude, latitude and their inter-correlation and statistical significance using corrplot R packages. Followingly, we merged all Chinese populations as one super-population and then defined all common lineages with a frequency larger than 0.01 or 0.05. We also used the corrplot R package to test the correlation between ADMIXTURE proportion and haplogroup frequency.

Declarations

Ethics approval and consent to participate

The Medical Ethics Committee of West China Hospital of Sichuan University approved this study. This study was conducted following the principles of the Helsinki Declaration.

Consent for publication

Not applicable.

Availability of data and materials

All haplogroup information was submitted in the supplementary materials. We followed the regulations of the Ministry of Science and Technology of the People's Republic of China. The raw genotype data required controlled access. Further requests for access to raw data can be directed to Guanglin He (Guanglinhescu@163.com) and Mengge Wang (menggewang2021@163.com).

Competing interests

The authors declare that they have no competing interests.

Funding

This work was supported by grants from the National Natural Science Foundation of China (82202078).

Authors' contributions

G.H., M.W. B.Z. and C.L. conceived and supervised the project. G.H. and M.W. collected the samples. K.L., K.Z., Y.H., G.H. and M.W. extracted the genomic DNA and performed the genome sequencing. G.H., M.W. and K.L. did variant calling. M.W., Y.H., K.L., H.Y., Z.W., S.D., L.W., H.Y., Q.S., J.Z., R.T., J.C., Y.S., X.L., C.W., H.S., Q.Y., L.H., L.Y., J.Y., S.N., Y.C., J.Y., K.Z., B.Z., C.L., G.H. performed population genetic analysis. G.H. and M.W. drafted the manuscript. G.H., M.W., B.Z. and C.L. revised the manuscript.

Acknowledgments

We thank all volunteers who participated in this project.

References

Alexander DH, Novembre J, Lange K. 2009. Fast model-based estimation of ancestry in unrelated individuals. Genome Res 19:1655-1664.

Bergstrom A, McCarthy SA, Hui R, Almarri MA, Ayub Q, Danecek P, Chen Y, Felkel S, Hallast P, Kamm J, et al. 2020. Insights into human genetic variation and population history from 929 diverse genomes. Science 367.

Byrska-Bishop M, Evani US, Zhao X, Basile AO, Abel HJ, Regier AA, Corvelo A, Clarke WE, Musunuri R, Nagulapalli K, et al. 2022. High-coverage whole-genome sequencing of the

expanded 1000 Genomes Project cohort including 602 trios. Cell 185:3426-3440 e3419.

Cao Y, Li L, Xu M, Feng Z, Sun X, Lu J, Xu Y, Du P, Wang T, Hu R, et al. 2020. The ChinaMAP analytics of deep whole genome sequences in 10,588 individuals. Cell Res 30:717-731.

Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. 2015. Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience 4:7.

Chen H, Lin R, Lu Y, Zhang R, Gao Y, He Y, Xu S. 2022. Tracing Bai-Yue Ancestry in Aboriginal Li People on Hainan Island. Mol Biol Evol 39.

Chen H, Lu Y, Lu D, Xu S. 2021. Y-LineageTracker: a high-throughput analysis framework for Y-chromosomal next-generation sequencing data. Bmc Bioinformatics 22:114.

Chen Z, Zhang Y, Fan A, Zhang Y, Wu Y, Zhao Q, Zhou Y, Zhou C, Bawudong M, Mao X, et al. 2011. Brief communication: Y-chromosome haplogroup analysis indicates that Chinese Tuvans share distinctive affinity with Siberian Tuvans. Am J Phys Anthropol 144:492-497.

Cheng S, Xu Z, Bian S, Chen X, Shi Y, Li Y, Duan Y, Liu Y, Lin J, Jiang Y, et al. 2023. The STROMICS genome study: deep whole-genome sequencing and analysis of 10K Chinese patients with ischemic stroke reveal complex genetic and phenotypic interplay. Cell Discov 9:75.

Choudhury A, Aron S, Botigue LR, Sengupta D, Botha G, Bensellak T, Wells G, Kumuthini J, Shriner D, Fakim YJ, et al. 2020. High-depth African genomes inform human migration and health. Nature 586:741-748.

Cong PK, Bai WY, Li JC, Yang MY, Khederzadeh S, Gai SR, Li N, Liu YH, Yu SH, Zhao WW, et al. 2022. Genomic analyses of 10,376 individuals in the Westlake BioBank for Chinese (WBBC) pilot project. Nat Commun 13:2939.

Cui Y, Li H, Ning C, Zhang Y, Chen L, Zhao X, Hagelberg E, Zhou H. 2013. Y Chromosome analysis of prehistoric human populations in the West Liao River Valley, Northeast China. BMC Evol Biol 13:216.

de Barros Damgaard P, Martiniano R, Kamm J, Moreno-Mayar JV, Kroonen G, Peyrot M, Barjamovic G, Rasmussen S, Zacho C, Baimukhanov N, et al. 2018. The first horse herders and the impact of early Bronze Age steppe expansions into Asia. Science 360.

Di Cristofaro J, Pennarun E, Mazieres S, Myres NM, Lin AA, Temori SA, Metspalu M, Metspalu E, Witzel M, King RJ, et al. 2013. Afghan Hindu Kush: where Eurasian sub-continent gene flows converge. PLoS One 8:e76748.

Feng Q, Lu Y, Ni X, Yuan K, Yang Y, Yang X, Liu C, Lou H, Ning Z, Wang Y, et al. 2017. Genetic History of Xinjiang's Uyghurs Suggests Bronze Age Multiple-Way Contacts in Eurasia. Mol Biol Evol 34:2572-2582.

Francalacci P, Morelli L, Angius A, Berutti R, Reinier F, Atzeni R, Pilu R, Busonero F, Maschio A, Zara I, et al. 2013. Low-pass DNA sequencing of 1200 Sardinians reconstructs European Y-chromosome phylogeny. Science 341:565-569.

Fuller DQ. 2007. Non-human genetics, agricultural origins and historical linguistics in South Asia. In. The evolution and history of human populations in south Asia: inter-disciplinary studies in archaeology, biological anthropology, linguistics and genetics: Springer. p. 393-443.

Grugni V, Battaglia V, Hooshiar Kashani B, Parolo S, Al-Zahery N, Achilli A, Olivieri A, Gandini F, Houshmand M, Sanati MH, et al. 2012. Ancient migratory events in the Middle East: new clues from the Y-chromosome variation of modern Iranians. PLoS One 7:e41252.

Hallast P, Agdzhoyan A, Balanovsky O, Xue Y, Tyler-Smith C. 2021. A Southeast Asian origin for present-day non-African human Y chromosomes. Hum Genet 140:299-307.

He G, Fan ZQ, Zou X, Deng X, Yeh HY, Wang Z, Liu J, Xu Q, Chen L, Deng XH, et al. 2022. Demographic model and biological adaptation inferred from the genome-wide single nucleotide polymorphism data reveal tripartite origins of southernmost Chinese Huis. American Journal of Biological Anthropology 180:488-505.

He G, Wang M, Li Y, Zou X, Yeh HY, Tang R, Yang X, Wang Z, Guo J, Luo T, et al. 2021. Fine-scale north-to-south genetic admixture profile in Shaanxi Han Chinese revealed by genome-wide demographic history reconstruction. Journal of Systematics and Evolution 60:955-972.

He G, Wang M, Miao L, Chen J, Zhao J, Sun Q, Duan S, Wang Z, Xu X, Sun Y, et al. 2023. Multiple founding paternal lineages inferred from the newly-developed 639-plex Y-SNP panel suggested the complex admixture and migration history of Chinese people. Hum Genomics 17:29.

He G, Wang M, Zou X, Chen P, Wang Z, Liu Y, Yao H, Wei LH, Tang R, Wang CC, et al. 2021. Peopling History of the Tibetan Plateau and Multiple Waves of Admixture of Tibetans Inferred From Both Ancient and Modern Genome-Wide Data. Front Genet 12:725243.

He GG, Yao H, Sun Q, Duan S, Tang R, Chen J, Wang Z, Sun Y, Li X, Wang S. 2023. Whole-genome sequencing of ethnolinguistic diverse northwestern Chinese Hexi Corridor people from the 10K_CPGDP project suggested the differentiated East-West genetic admixture along the Silk Road and their biological adaptations. bioRxiv:2023.2002. 2026.530053.

Helgason A, Einarsson AW, Guethmundsdottir VB, Sigurethsson A, Gunnarsdottir ED, Jagadeesan A, Ebenesersdottir SS, Kong A, Stefansson K. 2015. The Y-chromosome point mutation rate in humans. Nat Genet 47:453-457.

Hu K, Yan S, Liu K, Ning C, Wei L-H, Li S-L, Song B, Yu G, Chen F, Liu L-J. 2015. The

dichotomy structure of Y chromosome Haplogroup N. arXiv preprint arXiv:1504.06463.

Huang YZ, Pamjav H, Flegontov P, Stenzl V, Wen SQ, Tong XZ, Wang CC, Wang LX, Wei LH, Gao JY, et al. 2018. Dispersals of the Siberian Y-chromosome haplogroup Q in Eurasia. Mol Genet Genomics 293:107-117.

Huang YZ, Wei LH, Yan S, Wen SQ, Wang CC, Yang YJ, Wang LX, Lu Y, Zhang C, Xu SH, et al. 2018. Whole sequence analysis indicates a recent southern origin of Mongolian Y-chromosome C2c1a1a1-M407. Mol Genet Genomics 293:657-663.

Ilumae AM, Reidla M, Chukhryaeva M, Jarve M, Post H, Karmin M, Saag L, Agdzhoyan A, Kushniarevich A, Litvinov S, et al. 2016. Human Y Chromosome Haplogroup N: A Non-trivial Time-Resolved Phylogeography that Cuts across Language Families. Am J Hum Genet 99:163-173.

Jagadeesan A, Ebenesersdottir SS, Guethmundsdottir VB, Thordardottir EL, Moore KHS, Helgason A. 2021. HaploGrouper: a generalized approach to haplogroup classification. Bioinformatics 37:570-572.

Jama WMAJ. 2013. World Medical Association Declaration of Helsinki: ethical principles for medical research involving human subjects. 310:2191-2194.

Jeong C, Wang K, Wilkin S, Taylor WTT, Miller BK, Bemmann JH, Stahl R, Chiovelli C, Knolle F, Ulziibayar S, et al. 2020. A Dynamic 6,000-Year Genetic History of Eurasia's Eastern Steppe. Cell 183:890-904 e829.

Jobling MA, Tyler-Smith C. 2017. Human Y-chromosome variation in the genome-sequencing era. Nat Rev Genet 18:485-497.

Kayser M, Brauer S, Weiss G, Schiefenhovel W, Underhill P, Shen P, Oefner P, Tommaseo-Ponzetta M, Stoneking M. 2003. Reduced Y-chromosome, but not mitochondrial DNA,

diversity in human populations from West New Guinea. Am J Hum Genet 72:281-302.

Ke Y, Su B, Song X, Lu D, Chen L, Li H, Qi C, Marzuki S, Deka R, Underhill P, et al. 2001. African origin of modern humans in East Asia: a tale of 12,000 Y chromosomes. Science 292:1151-1153.

Kumar S, Stecher G, Li M, Knyaz C, Tamura K. 2018. MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms. Mol Biol Evol 35:1547-1549.

Kumar S, Stecher G, Tamura K. 2016. MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. Mol Biol Evol 33:1870-1874.

Kumar V, Wang W, Zhang J, Wang Y, Ruan Q, Yu J, Wu X, Hu X, Wu X, Guo W, et al. 2022. Bronze and Iron Age population movements underlie Xinjiang population history. Science 376:62-69.

Kutanan W, Kampuansai J, Srikummool M, Brunelli A, Ghirotto S, Arias L, Macholdt E, Hubner A, Schroder R, Stoneking M. 2019. Contrasting Paternal and Maternal Genetic Histories of Thai and Lao Populations. Mol Biol Evol 36:1490-1506.

Kutanan W, Shoocongdej R, Srikummool M, Hubner A, Suttipai T, Srithawong S, Kampuansai J, Stoneking M. 2020. Cultural variation impacts paternal and maternal genetic lineages of the Hmong-Mien and Sino-Tibetan groups from Thailand. Eur J Hum Genet 28:1563-1579.

Lang M, Liu H, Song F, Qiao X, Ye Y, Ren H, Li J, Huang J, Xie M, Chen S, et al. 2019. Forensic characteristics and genetic analysis of both 27 Y-STRs and 143 Y-SNPs in Eastern Han Chinese population. Forensic Sci Int Genet 42:e13-e20.

Leigh JW, Bryant D, Nakagawa S. 2015. popart: full-feature software for haplotype network construction. Methods in Ecology and Evolution 6:1110-1116.

Letunic I, Bork P. 2021. Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree

display and annotation. Nucleic Acids Res 49:W293-W296.

Lewis MP, Simons GF, Fennig CD. 2016. Languages of China: an Ethnologue country report. In: Dallas, TX: SIL International. http://www. ethnologue. com.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25:1754-1760.

Li H, Huang Y, Mustavich LF, Zhang F, Tan JZ, Wang LE, Qian J, Gao MH, Jin L. 2007. Y chromosomes of prehistoric people along the Yangtze River. Hum Genet 122:383-388.

Li J, Cai D, Zhang Y, Zhu H, Zhou H. 2020. Ancient DNA reveals two paternal lineages C2a1a1b1a/F3830 and C2b1b/F845 in past nomadic peoples distributed on the Mongolian Plateau. Am J Phys Anthropol 172:402-411.

Li J, Zeng W, Zhang Y, Ko AM, Li C, Zhu H, Fu Q, Zhou H. 2017. Ancient DNA reveals genetic connections between early Di-Qiang and Han Chinese. BMC Evol Biol 17:239.

Li YC, Ye WJ, Jiang CG, Zeng Z, Tian JY, Yang LQ, Liu KJ, Kong QP. 2019. River Valleys Shaped the Maternal Genetic Landscape of Han Chinese. Mol Biol Evol 36:1643-1652.

Liu BL, Ma PC, Wang CZ, Yan S, Yao HB, Li YL, Xie YM, Meng SL, Sun J, Cai YH, et al. 2021. Paternal origin of Tungusic-speaking populations: Insights from the updated phylogenetic tree of Y-chromosome haplogroup C2a-M86. Am J Hum Biol 33:e23462.

Liu D, Duong NT, Ton ND, Van Phong N, Pakendorf B, Van Hai N, Stoneking M. 2020. Extensive Ethnolinguistic Diversity in Vietnam Reflects Multiple Sources of Genetic Diversity. Mol Biol Evol 37:2503-2519.

Liu D, Ko AM, Stoneking M. 2023. The genomic diversity of Taiwanese Austronesian groups: Implications for the "Into- and Out-of-Taiwan" models. PNAS Nexus 2:pgad122.

Liu J, Jiang L, Zhao M, Du W, Wen Y, Li S, Zhang S, Fang F, Shen J, He G, et al. 2022.

Development and validation of a custom panel including 256 Y-SNPs for Chinese Y-chromosomal haplogroups dissection. Forensic Sci Int Genet 61:102786.

Liu Y, Xie J, Wang M, Liu C, Zhu J, Zou X, Li W, Wang L, Leng C, Xu Q, et al. 2021. Genomic Insights Into the Population History and Biological Adaptation of Southwestern Chinese Hmong-Mien People. Front Genet 12:815160.

Ma B, Chen J, Yang X, Bai J, Ouyang S, Mo X, Chen W, Wang CC, Hai X. 2021. The Genetic Structure and East-West Population Admixture in Northwest China Inferred From Genome-Wide Array Genotyping. Front Genet 12:795570.

Macholdt E, Arias L, Duong NT, Ton ND, Van Phong N, Schroder R, Pakendorf B, Van Hai N, Stoneking M. 2020. The paternal and maternal genetic history of Vietnamese populations. Eur J Hum Genet 28:636-645.

Mao X, Zhang H, Qiao S, Liu Y, Chang F, Xie P, Zhang M, Wang T, Li M, Cao P, et al. 2021. The deep population history of northern East Asia from the Late Pleistocene to the Holocene. Cell 184:3256-3266 e3213.

Myres NM, Rootsi S, Lin AA, Jarve M, King RJ, Kutuev I, Cabrera VM, Khusnutdinova EK, Pshenichnov A, Yunusbayev B, et al. 2011. A major Y-chromosome haplogroup R1b Holocene era founder effect in Central and Western Europe. Eur J Hum Genet 19:95-101.

Narasimhan VM, Patterson N, Moorjani P, Rohland N, Bernardos R, Mallick S, Lazaridis I, Nakatsuka N, Olalde I, Lipson M, et al. 2019. The formation of human populations in South and Central Asia. Science 365.

Nielsen R, Akey JM, Jakobsson M, Pritchard JK, Tishkoff S, Willerslev E. 2017. Tracing the peopling of the world through genomics. Nature 541:302-310.

Ning C, Fernandes D, Changmai P, Flegontova O, Yüncü E, Maier R, Altınışık NE, Kassian

AS, Krause J, Lalueza-Fox C. 2020. The genomic formation of First American ancestors in East and Northeast Asia. bioRxiv:2020.2010. 2012.336628.

Ning C, Li T, Wang K, Zhang F, Li T, Wu X, Gao S, Zhang Q, Zhang H, Hudson MJ, et al. 2020. Ancient genomes from northern China suggest links between subsistence changes and human migration. Nat Commun 11:2700.

Olalde I, Brace S, Allentoft ME, Armit I, Kristiansen K, Booth T, Rohland N, Mallick S, Szecsenyi-Nagy A, Mittnik A, et al. 2018. The Beaker phenomenon and the genomic transformation of northwest Europe. Nature 555:190-196.

Olson ND, Wagner J, Dwarshuis N, Miga KH, Sedlazeck FJ, Salit M, Zook JM. 2023. Variant calling and benchmarking in an era of complete human genome sequences. Nat Rev Genet 24:464-483.

Poznik GD, Henn BM, Yee MC, Sliwerska E, Euskirchen GM, Lin AA, Snyder M, Quintana-Murci L, Kidd JM, Underhill PA, et al. 2013. Sequencing Y chromosomes resolves discrepancy in time to common ancestor of males versus females. Science 341:562-565.

Poznik GD, Xue Y, Mendez FL, Willems TF, Massaia A, Wilson Sayres MA, Ayub Q, McCarthy SA, Narechania A, Kashin S, et al. 2016. Punctuated bursts in human male demography inferred from 1,244 worldwide Y-chromosome sequences. Nat Genet 48:593-599.

Qi X, Cui C, Peng Y, Zhang X, Yang Z, Zhong H, Zhang H, Xiang K, Cao X, Wang Y, et al. 2013. Genetic evidence of paleolithic colonization and neolithic expansion of modern humans on the tibetan plateau. Mol Biol Evol 30:1761-1778.

Raghavan M, Skoglund P, Graf KE, Metspalu M, Albrechtsen A, Moltke I, Rasmussen S, Stafford TW, Jr., Orlando L, Metspalu E, et al. 2014. Upper Palaeolithic Siberian genome

reveals dual ancestry of Native Americans. Nature 505:87-91.

Ralf A, Montiel Gonzalez D, Zhong K, Kayser M. 2018. Yleaf: Software for Human Y-Chromosomal Haplogroup Inference from Next-Generation Sequencing Data. Mol Biol Evol 35:1291-1294.

Ralf A, van Oven M, Zhong K, Kayser M. 2015. Simultaneous analysis of hundreds of Y-chromosomal SNPs for high-resolution paternal lineage classification using targeted semiconductor sequencing. Hum Mutat 36:151-159.

Rootsi S, Zhivotovsky LA, Baldovic M, Kayser M, Kutuev IA, Khusainova R, Bermisheva MA, Gubina M, Fedorova SA, Ilumae AM, et al. 2007. A counter-clockwise northern route of the Y-chromosome haplogroup N from Southeast Asia towards Europe. Eur J Hum Genet 15:204-211.

Semino O, Passarino G, Oefner PJ, Lin AA, Arbuzova S, Beckman LE, Benedictis GD, Francalacci P, Kouvatsi A, Limborska S. 2000. The genetic legacy of Paleolithic Homo sapiens sapiens in extant Europeans: AY chromosome perspective. Science 290:1155-1159.

Shi H, Dong YL, Wen B, Xiao CJ, Underhill PA, Shen PD, Chakraborty R, Jin L, Su B. 2005. Y-chromosome evidence of southern origin of the East Asian-specific haplogroup O3-M122. Am J Hum Genet 77:408-419.

Shi H, Qi X, Zhong H, Peng Y, Zhang X, Ma RZ, Su B. 2013. Genetic evidence of an East Asian origin and paleolithic northward migration of Y-chromosome haplogroup N. PLoS One 8:e66102.

Shi H, Zhong H, Peng Y, Dong YL, Qi XB, Zhang F, Liu LF, Tan SJ, Ma RZ, Xiao CJ, et al. 2008. Y chromosome evidence of earliest modern human settlement in East Asia and multiple origins of Tibetan and Japanese populations. BMC Biol 6:45.

Shou WH, Qiao EF, Wei CY, Dong YL, Tan SJ, Shi H, Tang WR, Xiao CJ. 2010. Y-chromosome distributions among populations in Northwest China identify significant contribution from Central Asian pastoralists and lesser influence of western Eurasians. J Hum Genet 55:314-322.

Sirugo G, Williams SM, Tishkoff SA. 2019. The Missing Diversity in Human Genetic Studies. Cell 177:26-31.

Song M, Wang Z, Lyu Q, Ying J, Wu Q, Jiang L, Wang F, Zhou Y, Song F, Luo H. 2022. Paternal genetic structure of the Qiang ethnic group in China revealed by high-resolution Y-chromosome STRs and SNPs. Forensic Science International: Genetics 61:102774.

Song M, Wang Z, Zhang Y, Zhao C, Lang M, Xie M, Qian X, Wang M, Hou Y. 2019. Forensic characteristics and phylogenetic analysis of both Y-STR and Y-SNP in the Li and Han ethnic groups from Hainan Island of China. Forensic Sci Int Genet 39:e14-e20.

Su B, Xiao J, Underhill P, Deka R, Zhang W, Akey J, Huang W, Shen D, Lu D, Luo J, et al. 1999. Y-Chromosome evidence for a northward migration of modern humans into Eastern Asia during the last Ice Age. Am J Hum Genet 65:1718-1724.

Sun J, Li YX, Ma PC, Yan S, Cheng HZ, Fan ZQ, Deng XH, Ru K, Wang CC, Chen G, et al. 2021. Shared paternal ancestry of Han, Tai-Kadai-speaking, and Austronesian-speaking populations as revealed by the high resolution phylogeny of O1a-M119 and distribution of its sub-lineages within China. Am J Phys Anthropol 174:686-700.

Sun N, Ma PC, Yan S, Wen SQ, Sun C, Du PX, Cheng HZ, Deng XH, Wang CC, Wei LH. 2019. Phylogeography of Y-chromosome haplogroup Q1a1a-M120, a paternal lineage connecting populations in Siberia and East Asia. Ann Hum Biol 46:261-266.

Tao R, Li M, Chai S, Xia R, Qu Y, Yuan C, Yang G, Dong X, Bian Y, Zhang S, et al. 2023.

Developmental validation of a 381 Y-chromosome SNP panel for haplogroup analysis in the Chinese populations. Forensic Sci Int Genet 62:102803.

Trejaut JA, Poloni ES, Yen JC, Lai YH, Loo JH, Lee CL, He CL, Lin M. 2014. Taiwan Y-chromosomal DNA variation and its relationship with Island Southeast Asia. BMC Genet 15:77.

Van Geystelen A, Decorte R, Larmuseau MH. 2013. AMY-tree: an algorithm to use whole genome SNP calling for Y chromosomal phylogenetic applications. BMC Genomics 14:101.

Wang C-C, Yan S, Qin Z-D, Lu Y, Ding Q-L, Wei L-H, Li S-L, Yang Y-J, Jin L, Li H. 2013. Late Neolithic expansion of ancient Chinese revealed by Y chromosome haplogroup O3a1c-002611. Journal of Systematics and Evolution 51:280-286.

Wang CC, Li H. 2013. Inferring human history in East Asia from Y chromosomes. Investig Genet 4:11.

Wang CC, Wang LX, Shrestha R, Zhang M, Huang XY, Hu K, Jin L, Li H. 2014. Genetic structure of Qiangic populations residing in the western Sichuan corridor. PLoS One 9:e103772.

Wang CC, Yeh HY, Popov AN, Zhang HQ, Matsumura H, Sirak K, Cheronet O, Kovalev A, Rohland N, Kim AM, et al. 2021. Genomic insights into the formation of human populations in East Asia. Nature 591:413-419.

Wang F, Song F, Song M, Li J, Xie M, Hou Y. 2021. Genetic reconstruction and phylogenetic analysis by 193 Y-SNPs and 27 Y-STRs in a Chinese Yi ethnic group. Electrophoresis 42:1480-1487.

Wang F, Song F, Song M, Luo H, Hou Y. 2022. Genetic structure and paternal admixture of the modern Chinese Zhuang population based on 37 Y-STRs and 233 Y-SNPs. Forensic Sci

Int Genet 58:102681.

Wang J, Yang L, Duan S, Sun Q, Li Y, Wu J, Wu W, Wang Z, Liu Y, Tang R, et al. 2023. Genome-wide allele and haplotype-sharing patterns suggested one unique Hmong-Mein-related lineage and biological adaptation history in Southwest China. Hum Genomics 17:3.

Wang LX, Lu Y, Zhang C, Wei LH, Yan S, Huang YZ, Wang CC, Mallick S, Wen SQ, Jin L, et al. 2018. Reconstruction of Y-chromosome phylogeny reveals two neolithic expansions of Tibeto-Burman populations. Mol Genet Genomics 293:1293-1300.

Wang M, He G, Zou X, Liu J, Ye Z, Ming T, Du W, Wang Z, Hou Y. 2021. Genetic insights into the paternal admixture history of Chinese Mongolians via high-resolution customized Y-SNP SNaPshot panels. Forensic Sci Int Genet 54:102565.

Wang M, Wang Z, He G, Liu J, Wang S, Qian X, Lang M, Li J, Xie M, Li C, et al. 2019. Developmental validation of a custom panel including 165 Y-SNPs for Chinese Y-chromosomal haplogroups dissection using the ion S5 XL system. Forensic Sci Int Genet 38:70-76.

Wang MG, He GL, Zou X, Chen PY, Wang Z, Tang RK, Yang XM, Chen J, Yang MQ, Li YX, et al. 2022. Reconstructing the genetic admixture history of Tai-Kadai and Sinitic people: Insights from genome-wide SNP data from South China. Journal of Systematics and Evolution 61:157-178.

Wang T, Wang W, Xie G, Li Z, Fan X, Yang Q, Wu X, Cao P, Liu Y, Yang R, et al. 2021. Human population history at the crossroads of East and Southeast Asia since 11,000 years ago. Cell 184:3829-3841 e3821.

Wei LH, Huang YZ, Yan S, Wen SQ, Wang LX, Du PX, Yao DL, Li SL, Yang YJ, Jin L, et al. 2017. Phylogeny of Y-chromosome haplogroup C3b-F1756, an important paternal lineage in

Altaic-speaking populations. J Hum Genet 62:915-918.

Wei LH, Yan S, Lu Y, Wen SQ, Huang YZ, Wang LX, Li SL, Yang YJ, Wang XF, Zhang C, et al. 2018. Whole-sequence analysis indicates that the Y chromosome C2*-Star Cluster traces back to ordinary Mongols, rather than Genghis Khan. Eur J Hum Genet 26:230-237.

Wei W, Ayub Q, Chen Y, McCarthy S, Hou Y, Carbone I, Xue Y, Tyler-Smith C. 2013. A calibrated human Y-chromosomal phylogeny based on resequencing. Genome Res 23:388-395.

Wu Q, Cheng HZ, Sun N, Ma PC, Sun J, Yao HB, Xie YM, Li YL, Meng SL, Zhabagin M, et al. 2020. Phylogenetic analysis of the Y-chromosome haplogroup C2b-F1067, a dominant paternal lineage in Eastern Eurasia. J Hum Genet 65:823-829.

Xia Z-Y, Yan S, Wang C-C, Zheng H-X, Zhang F, Liu Y-C, Yu G, Yu B-X, Shu L-L, Jin L. 2019. Inland-coastal bifurcation of southern East Asians revealed by Hmong-Mien genomic history.

Xie M, Song F, Li J, Lang M, Luo H, Wang Z, Wu J, Li C, Tian C, Wang W, et al. 2019. Genetic substructure and forensic characteristics of Chinese Hui populations using 157 Y-SNPs and 27 Y-STRs. Forensic Sci Int Genet 41:11-18.

Xue Y, Wang Q, Long Q, Ng BL, Swerdlow H, Burton J, Skuce C, Taylor R, Abdellah Z, Zhao Y, et al. 2009. Human Y chromosome base-substitution mutation rate measured by direct sequencing in a deep-rooting pedigree. Curr Biol 19:1453-1457.

Yan S, Wang CC, Li H, Li SL, Jin L, Genographic C. 2011. An updated tree of Y-chromosome Haplogroup O and revised phylogenetic positions of mutations P164 and PK4. Eur J Hum Genet 19:1013-1015.

Yan S, Wang CC, Zheng HX, Wang W, Qin ZD, Wei LH, Wang Y, Pan XD, Fu WQ, He YG, et al. 2014. Y chromosomes of 40% Chinese descend from three Neolithic super-grandfathers.

PLoS One 9:e105691.

Yang MA, Fan X, Sun B, Chen C, Lang J, Ko YC, Tsang CH, Chiu H, Wang T, Bao Q, et al. 2020. Ancient DNA indicates human population shifts and admixture in northern and southern China. Science 369:282-288.

Yao X, Tang S, Bian B, Wu X, Chen G, Wang CC. 2017. Improved phylogenetic resolution for Y-chromosome Haplogroup O2a1c-002611. Sci Rep 7:1146.

Yu HX, Ao C, Wang XP, Zhang XP, Sun J, Li H, Liu KJ, Wei LH. 2023. The impacts of bronze age in the gene pool of Chinese: Insights from phylogeographics of Y-chromosomal haplogroup N1a2a-F1101. Front Genet 14:1139722.

Zerjal T, Xue Y, Bertorelle G, Wells RS, Bao W, Zhu S, Qamar R, Ayub Q, Mohyuddin A, Fu S, et al. 2003. The genetic legacy of the Mongols. Am J Hum Genet 72:717-721.

Zhabagin M, Wei LH, Sabitov Z, Ma PC, Sun J, Dyussenova Z, Balanovska E, Li H, Ramankulov Y. 2022. Ancient Components and Recent Expansion in the Eurasian Heartland: Insights into the Revised Phylogeny of Y-Chromosomes from Central Asia. Genes (Basel) 13.

Zhang F, Ning C, Scott A, Fu Q, Bjorn R, Li W, Wei D, Wang W, Fan L, Abuduresule I, et al. 2021. The genomic origins of the Bronze Age Tarim Basin mummies. Nature 599:256-261.

Zhang P, Luo H, Li Y, Wang Y, Wang J, Zheng Y, Niu Y, Shi Y, Zhou H, Song T, et al. 2021. NyuWa Genome resource: A deep whole-genome sequencing-based variation profile and reference panel for the Chinese population. Cell Rep 37:110017.

Zhang X, Kampuansai J, Qi X, Yan S, Yang Z, Serey B, Sovannary T, Bunnath L, Aun HS, Samnom H, et al. 2014. An updated phylogeny of the human Y-chromosome lineage O2a-M95 with novel SNPs. PLoS One 9:e101020.

Zhang Y, Lei X, Chen H, Zhou H, Huang S. 2018. Ancient DNAs and the Neolithic Chinese

super-grandfather Y haplotypes. bioRxiv:487918.

Zhang Y, Wu X, Li J, Li H, Zhao Y, Zhou H. 2018. The Y-chromosome haplogroup C3*-F3918, likely attributed to the Mongol Empire, can be traced to a 2500-year-old nomadic group. J Hum Genet 63:231-238.

Zhang Z, Zhang Y, Wang Y, Zhao Z, Yang M, Zhang L, Zhou B, Xu B, Zhang H, Chen T, et al. 2022. The Tibetan-Yi region is both a corridor and a barrier for human gene flow. Cell Rep 39:110720.

Zhao YB, Zhang Y, Li HJ, Cui YQ, Zhu H, Zhou H. 2014. Ancient DNA evidence reveals that the Y chromosome haplogroup Q1a1 admixed into the Han Chinese 3,000 years ago. Am J Hum Biol 26:813-821.

Zhong H, Shi H, Qi XB, Duan ZY, Tan PP, Jin L, Su B, Ma RZ. 2011. Extended Y chromosome investigation suggests postglacial migrations of modern humans into East Asia via the northern route. Mol Biol Evol 28:717-727.

Zhong H, Shi H, Qi XB, Xiao CJ, Jin L, Ma RZ, Su B. 2010. Global distribution of Y-chromosome haplogroup C reveals the prehistoric migration routes of African exodus and early settlement in East Asia. J Hum Genet 55:428-435.

Fig. 1. Geographical position, sample size and phylogenetic features of newly-generated 919 targeted sequences. (**A**) A map of East Asia showed the basic information of 919 individuals from 57 geographically or ethnically different populations. The circle size in the map denoted the sample size of individual populations. The colored provinces indicated sampling places, and the color of the province showed the total sample size from these geographical regions. Ancient subsistence strategies (pastoralists, fisher-hunter-gatherers and farmers) from western Eurasia, the Mongolian Plateau and Chinese agriculture original centers (Yellow and Yangtze River Basin) were also denoted. (**B**) Network relationships among 919 haplotypes were inferred based on the median-joining network algorithms. Different colors of the circle showed the geographical origin of one focused haplogroup, and different branches were labeled their main Y-lineages and their possible related ancestral East Asians. (**C**) Phylogeny of Chinese minority ethnic groups denoted the different lineages. Different colors denoted the different older upper-stream lineages.

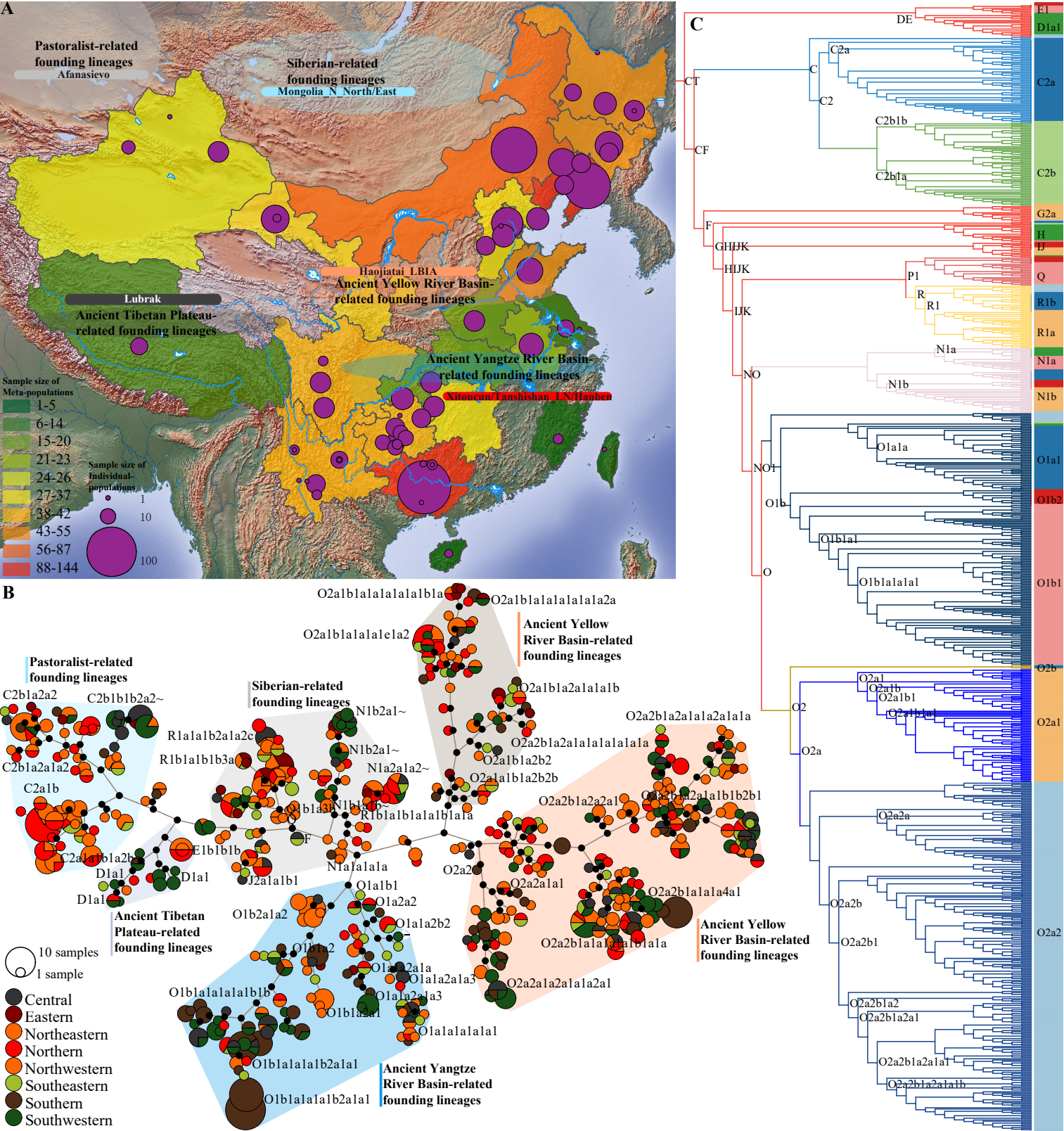
Fig. 2. Population genetic structure among 130 modern and four ancient East Asians. (**A**) Principal Component Analysis (PCA) and Multidimensional Scaling plots (MDS) based on the top two components showed the genetic similarities and differences between 134 populations, including 13,886 individuals. (**B-C**) Heatmap showed genetic affinity among Altaic- and Southern Chinese indigenous and Southeast Asian populations. The genetic Fst matrix among Sino-Tibetan-speaking populations was presented in **Fig. S3**. (**D**) Phylogenetic relationships reconstructed based on the Fst matrix and frequency distribution of major Y-chromosome lineages.

Fig. 3. The frequency spectrum of Chinese-dominant Y-chromosome lineages among ancient

Eurasians and modern ethnolinguistically diverse East Asians and Southeast Asians. (**A**) The geographical position of 1284 ancient individuals carrying twelve Y-chromosome lineages is interesting in the present work. Different colors of circles indicated different haplogroups, and the circle size denoted the number of interested haplogroups. (**B-C**) Haplogroup frequency of western-origin and Siberian hunter-gatherer-related lineages among eastern Eurasian populations. Optimized hot spot analysis suggested the geographical origin of the focused lineages. (**D-E**) Haplogroup frequency of sublineages derived from the early East Asian-related D, ancient northern East Asian millet farmer-related O2, and ancient southern East Asian rice farmer-related O1. The hot red color suggested the high-frequency or phylogeographical regions of the studied lineages. The frequency distribution of other sublineages was presented in detail in **Figs. S4-7**.

Fig.4 Correlation between frequency of the Chinese-dominant Y-chromosome lineages and other geographical and genetic features. (**A**) The correlation of frequency of Y-chromosome lineages and geographical features of latitude and longitude, as well as with the top two components of the PCA. (**B**) The correlation between the frequency of Y-chromosome lineages and Fst matrix among eastern Eurasian populations. (**C**) The correlation efficient matrix among the pairwise pairs of haplogroup frequency. Asterisk-labeled tested pairs showed that the Spearman correlation is statistically significant. Three asterisks showed p values less than 0.001; two asterisks showed p values ranging from 0.001 to 0.1, and one asterisk indicated the p values ranged from 0.1 to 0.5. The blue color indicated the positive correlation, and the red color indicated the negative correlation.

Fig.5 Correlation between autosome-estimated ancestral proportion and Y-chromosome dominant lineages. (**A**) Model-based ADMIXTURE results among modern and ancient East Asians with predefined ancestral sources ranging from 2 to 10. The six-way admixture model with the less cross-validation error was the best-fitted model. (**B-G**) Admixture proportion distribution among different Chinese populations. The red color denoted the highest proportion of one targeted ancestral component. (**H-M**) Scatter plots showed statistically positive correlations between autosome-based ancestral proportion and putatively population-specific founding lineages. (**N**) The correlation between autosome-based ADMIXTURE estimates of ancestral proportion and frequency of Y-chromosome lineages. The correlations between proportions of different ancestral sources were manually removed, and their initial clustering positions were labeled with arrows. (**O**) Venn diagram showed shared and specific lineages among different autosome-based putative ancestral source-related lineages.


A
Pastoralist-related founding lineages
Afanasievo
Siberian-related founding lineages
Mongolia_N_North/East
Haojiatai_LBIA
Ancient Yellow River Basin-related founding lineages
Lubrak
Ancient Tibetan Plateau-related founding lineages
Ancient Yangtze River Basin-related founding lineages
Xitoucun/Tanshishan_LN/Hanben
Sample size of Meta-populations
1-5
6-14
15-20
21-23
24-26
27-37
38-42
43-55
56-87
88-144
Sample size of individual-populations
1
10
100
B
Pastoralist-related founding lineages
Siberian-related founding lineages
Ancient Yellow River Basin-related founding lineages
Ancient Tibetan Plateau-related founding lineages
Ancient Yangtze River Basin-related founding lineages
10 samples
1 sample
Central
Eastern
Northeastern
Northern
Northwestern
Southeastern
Southern
Southwestern
C
DE
CT
C
C2a
C2
C2b1b
CF
C2b1a
F
GHIJK
HIJK
IJK
P1
R
R1
NO
N1a
N1b
NO1
O1a1a
O1b
O1b1a1
O
O1b1a1a1a1
O2
O2a
O2a1
O2a1a
O2a1a1
O2a1a1b
O2a1a1b1
O2a1a1b1a1
O2a2a
O2a2b
O2a2b1
O2a2b1a2
O2a2b1a2a1
O2a2b1a2a1a1
O2a2b1a2a1a1b
E1
D1a1
C2a
C2b
G2a
H
IJ
Q
R1b
R1a
N1a
N1b
O1a1
O1b2
O1b1
O2b
O2a1
O2a2

A

Ancient DNA
TransEurasian
Sinitic_South
Hmong_Mien/Austroasiatic
Sinitic_Hui
Tibeto_Burman
Hmong_Mien/Austroasiatic
Sinitic_Hui
Tibeto_Burman
Austronesian/Tai_Kadai
Sinitic_North
Austronesian/Tai_Kadai
Sinitic_North
Ancient DNA
TransEurasian
Sinitic_South

B

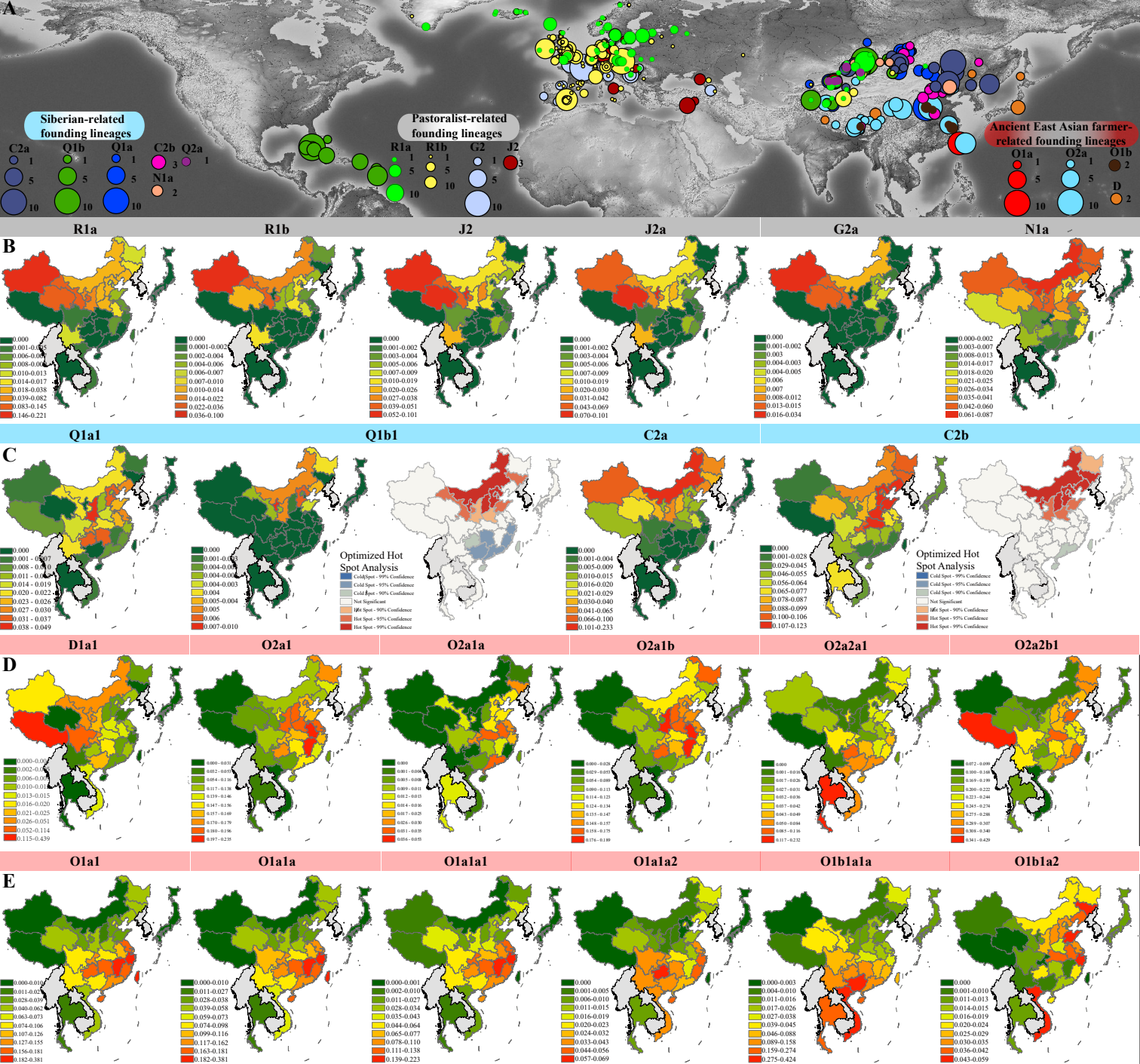
Group
Korean_Heilongjiang
Korean_Jilin
Manchu_Hebei
Daur_InnerMongolia
Manchu_Heilongjiang
Manchu_Liaoning
Manchu_Shandong
Manchu_Beijing
Hezhen_Heilongjiang
Daur_Heilongjiang
Manchu_Jilin
Mongolian_Liaoning
Mongolian_InnerMongolia
Kazakh_Xinjiang
Uyghur_Xinjiang
Xibe_Liaoning
Xibe_Xinjiang
Oroqen_InnerMongolia
Japanese_Japan
Japanese_Tokyo
Japonic
Koreanic
Mongolic
Tungusic
Turkic
Frequency
Colored ranges
Ancient_China
Austroasiatic
Austronesian
Tai_Kadai
Tibeto_Burman
Tungusic
Turkic
Hmong_Mien
Japonic
Koreanic
Mongolic
Sinitic_Hui
Sinitic_North
Sinitic_South

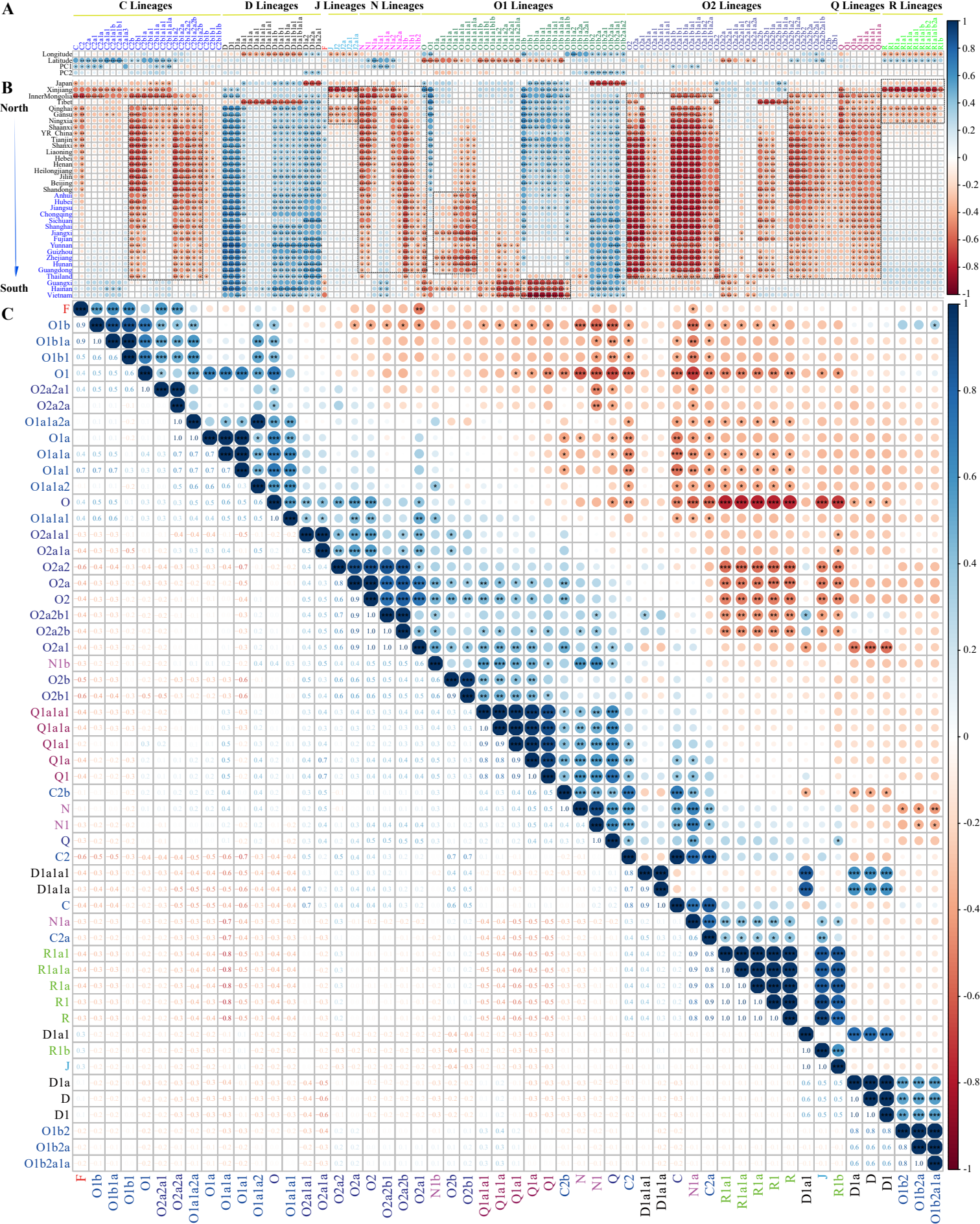
C

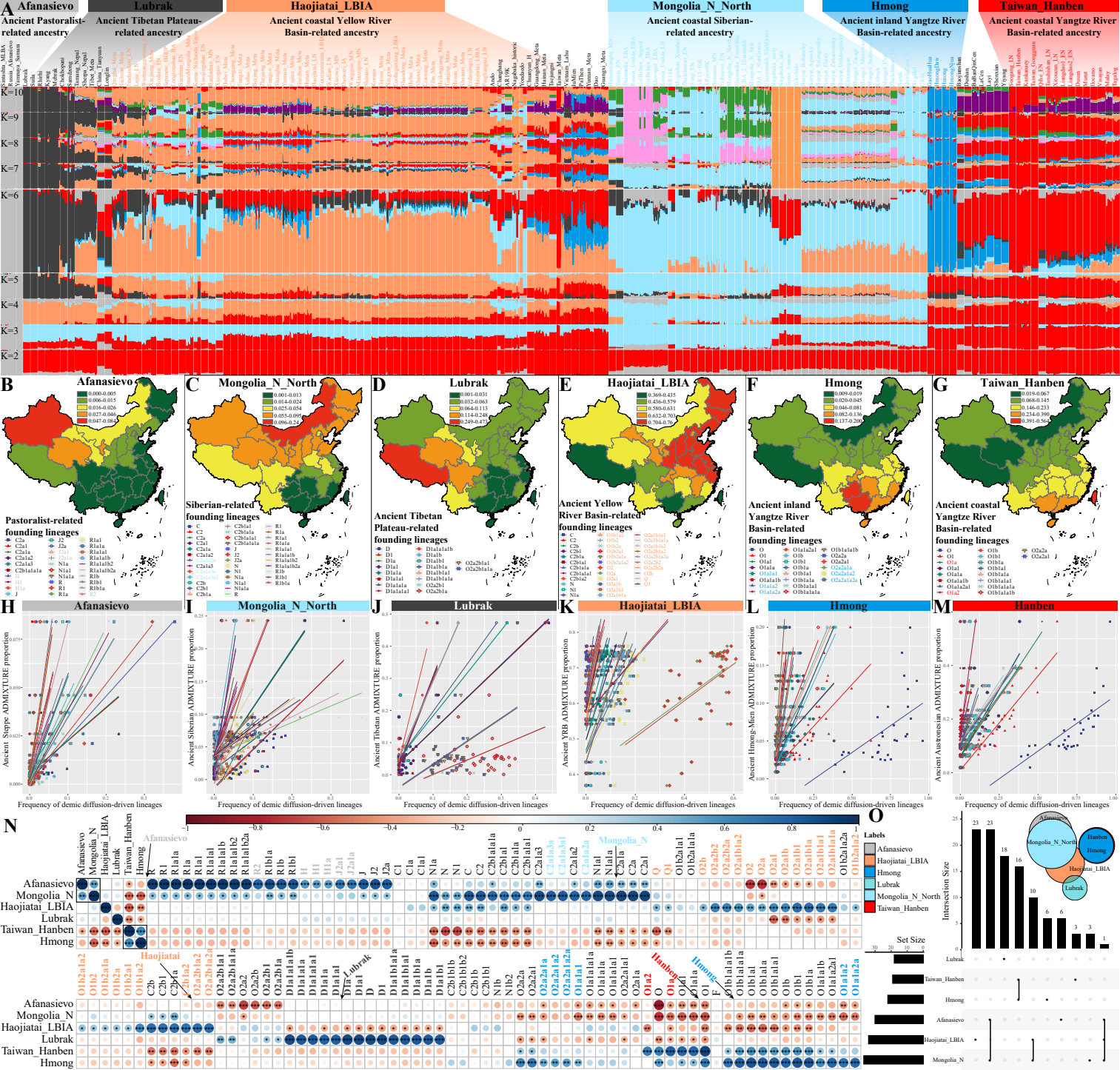
Group
Rukai_Taiwan
Thao_Taiwan
Saisiyat_Taiwan
Taroko_Taiwan
Tsou_Taiwan
Atayal_Taiwan
Mang_Lai_Chau
Bunun_Taiwan
She_Fujian
Hmong
Pathen_Ha_Giang
Shan
Dong_Guizhou
Gelao_Guizhou
Yami_Taiwan
Paiwan_Taiwan
Puyuma_Taiwan
Kategalan_Taiwan
Amis_Taiwan
Siraya_Taiwan
Lachi_Ha_Giang
LaoIsan
Ede_Dak_Lak
Colao_Ha_Giang
Yao_Guangxi
Nung_Ha_Giang
Giarai_Gia_Lai
Thai
Li_Hainan
Cambodian
Phutai
Tay
Miao_Guizhou
Dao_Ha_Giang
Dai_Yunnan
Zhuang_Guangxi
Bouyei_Guizhou
Kinh
Miao_Hunan
Kinh_HoChiMinh
IuMien
Austroasiatic
Austronesian
Hmong_Mien
Tai_Kadai

D

Tai-Kadai-related O1b lineages
Han Chinese-related O2a1 lineages
Austronesian-related O1a1 lineages
Altaic-related C2a/C2b lineages
Hmong-Mien-related O2a2 lineages


A
Siberian-related founding lineages
C2a
Q1b
Q1a
C2b
Q2a
N1a
Pastoralist-related founding lineages
R1a
R1b
G2
J2
Ancient East Asian farmer-related founding lineages
O1a
O2a
O1b
D
B
R1a
R1b
J2
J2a
G2a
N1a
Q1a1
Q1b1
C2a
C2b
C
Optimized Hot Spot Analysis
Cold Spot - 99% Confidence
Cold Spot - 95% Confidence
Cold Spot - 90% Confidence
Not Significant
Hot Spot - 90% Confidence
Hot Spot - 95% Confidence
Hot Spot - 99% Confidence
D1a1
O2a1
O2a1a
O2a1b
O2a2a1
O2a2b1
D
O1a1
O1a1a
O1a1a1
O1a1a2
O1b1a1a
O1b1a2
E


A
C Lineages
D Lineages
J Lineages
N Lineages
O1 Lineages
O2 Lineages
Q Lineages
R Lineages
Longitude
Latitude
PC1
PC2
B
North
South
Japan
Xinjiang
InnerMongolia
Tibet
Qinghai
Gansu
Ningxia
Shaanxi
YR. China
Tianjin
Shanxi
Liaoning
Hebei
Henan
Heilongjiang
Jilin
Beijing
Shandong
Anhui
Hubei
Jiangsu
Chongqing
Sichuan
Shanghai
Jiangxi
Fujian
Yunnan
Guizhou
Zhejiang
Hunan
Guangdong
Thailand
Guangxi
Hainan
Vietnam
C
F
O1b
O1b1a
O1b1
O1
O2a2a1
O2a2a
O1a1a2a
O1a
O1a1a
O1a1
O1a1a2
O
O1a1a1
O2a1a1
O2a1a
O2a2
O2a
O2
O2a2b1
O2a2b
O2a1
N1b
O2b
O2b1
Q1a1a1
Q1a1a
Q1a1
Q1a
Q1
C2b
N
N1
Q
C2
D1a1a1
D1a1a
C
N1a
C2a
R1a1
R1a1a
R1a
R1
R
D1a1
J
R1b
D1a
D
D1
O1b2
O1b2a
O1b2a1a


A
Afanasievo
Ancient Pastoralist-related ancestry
Lubrak
Ancient Tibetan Plateau-related ancestry
Haojiatai_LBIA
Ancient coastal Yellow River Basin-related ancestry
Mongolia_N_North
Ancient coastal Siberian-related ancestry
Hmong
Ancient inland Yangtze River Basin-related ancestry
Taiwan_Hanben
Ancient coastal Yangtze River Basin-related ancestry
K=10
K=9
K=8
K=7
K=6
K=5
K=4
K=3
K=2
B
Afanasievo
Pastoralist-related founding lineages
C
Mongolia_N_North
Siberian-related founding lineages
D
Lubrak
Ancient Tibetan Plateau-related founding lineages
E
Haojiatai_LBIA
Ancient Yellow River Basin-related founding lineages
F
Hmong
Ancient inland Yangtze River Basin-related founding lineages
G
Taiwan_Hanben
Ancient coastal Yangtze River Basin-related founding lineages
H
Afanasievo
Ancient Steppe ADMIXTURE proportion
I
Mongolia_N_North
Ancient Siberian ADMIXTURE proportion
J
Lubrak
Ancient Tibetan ADMIXTURE proportion
K
Haojiatai_LBIA
Ancient YRB ADMIXTURE proportion
L
Hmong
Ancient Hmong-Mien ADMIXTURE proportion
M
Hanben
Ancient Austronesian ADMIXTURE proportion
Frequency of demic diffusion-driven lineages
N
O
Labels
Afanasievo
Haojiatai_LBIA
Hmong
Lubrak
Mongolia_N_North
Taiwan_Hanben
Intersection Size
Set Size