https://doi.org/10.1186/s40246-023-00476-6

Human Genomics

RESEARCH

Open Access

Multiple founding paternal lineages inferred from the newly-developed 639-plex Y-SNP panel suggested the complex admixture and migration history of Chinese people

Guanglin He[1*†], Mengge Wang[2†], Lei Miao[3†], Jing Chen[4], Jie Zhao[3], Qiuxia Sun[1,5], Shuhan Duan[1,6], Zhiyong Wang[1,7], Xiaofei Xu[1], Yuntao Sun[1,8], Yan Liu[1,6], Jing Liu[8], Zheng Wang[8], Lanhai Wei[9], Chao Liu[2,10], Jian Ye[3*] and Le Wang[3*]

Abstract

Background Non-recombining regions of the Y-chromosome recorded the evolutionary traces of male human populations and are inherited haplotype-dependently and male-specifically. Recent whole Y-chromosome sequencing studies have identified previously unrecognized population divergence, expansion and admixture processes, which promotes a better understanding and application of the observed patterns of Y-chromosome genetic diversity.

Results Here, we developed one highest-resolution Y-chromosome single nucleotide polymorphism (Y-SNP) panel targeted for uniparental genealogy reconstruction and paternal biogeographical ancestry inference, which included 639 phylogenetically informative SNPs. We genotyped these loci in 1033 Chinese male individuals from 33 ethnolinguistically diverse populations and identified 256 terminal Y-chromosomal lineages with frequency ranging from 0.0010 (singleton) to 0.0687. We identified six dominant common founding lineages associated with different ethnolinguistic backgrounds, which included O2a2b1a1a1a1a1a1a1-M6539, O2a1b1a1a1a1a1a1-F17, O2a2b1a1a1a1a1b1a1b-MF15397, O2a2b2a1b1-A16609, O1b1a1a1a1b2a1a1-F2517, and O2a2b1a1a1a1a1a1-F155. The AMOVA and nucleotide diversity estimates revealed considerable differences and high genetic diversity among ethnolinguistically different populations. We constructed one representative phylogenetic tree among 33 studied populations based on the haplogroup frequency spectrum and sequence variations. Clustering patterns in principal component analysis and multidimensional scaling results showed a genetic differentiation between Tai-Kadai-speaking Li, Mongolic-speaking Mongolian, and other Sinitic-speaking Han Chinese populations. Phylogenetic topology inferred from the BEAST and Network relationships reconstructed from the popART further showed the founding lineages from culturally/linguistically diverse populations, such as C2a/C2b was dominant in Mongolian people and O1a/

[†] Guanglin He, Mengge Wang and Lei Miao these authors contributed equally and are considered as co-first authors

*Correspondence:
Guanglin He
guanglinhescu@163.com
Jian Ye
yejian77@126.com
Le Wang
wangle_02@163.com
Full list of author information is available at the end of the article

O1b was dominant in island Li people. We also identified many lineages shared by more than two ethnolinguistically different populations with a high proportion, suggesting their extensive admixture and migration history.

Conclusions Our findings indicated that our developed high-resolution Y-SNP panel included major dominant Y-lineages of Chinese populations from different ethnic groups and geographical regions, which can be used as the primary and powerful tool for forensic practice. We should emphasize the necessity and importance of whole sequencing of more ethnolinguistically different populations, which can help identify more unrecognized population-specific variations for the promotion of Y-chromosome-based forensic applications.

Keywords Phylogenetic tree, Y-SNPs, Population structure, Founding lineage, Network relationship, Biogeographical ancestry inference

Introduction

Human genomics studies in the whole-genome sequencing era have updated our understanding of the patterns of genetic diversity among ethnolinguistically diverse worldwide populations, fine-scale population structure, variation spectrum of various kinds of genetic variations of single nucleotide variations (SNV), structural variations (SV) and complex mobile elements [1–3]. The properties of the non-recombining region of the human Y-chromosome (NRY), that is, male specificity, haploidy, and absence of crossing over, make its genetic variations a powerful tool in evolutionary studies and forensic investigations, especially in cases where standard autosomal short tandem repeat (STR) profiling is not informative [4, 5]. Haplotypes composed of Y-STRs or Y-SNPs have been applied to characterize the paternal lineages of unknown male contributors or infer paternal biogeographical ancestry [5–7]. Y-SNPs define stable haplotypes as haplogroups, which could be used to construct robust phylogeny. A large body of research has been dedicated to identifying novel Y-chromosomal genetic variations, detailing the phylogenetic tree, and characterizing the patterns of differentiated paternal lineages [7–9]. Nowadays, the advances and applications of next-generation sequencing (NGS) technology provide unbiased ascertainment of Y-SNPs, leading to the construction of detailed phylogenies in which branch lengths are proportional to divergence times and enabling the estimates of time to the most recent common ancestor (TMRCA) of branch nodes [7, 10, 11]. The Y-chromosomal molecular clock built a link between genetic diversity and human migration and admixture history, which could be used to estimate the time when a lineage originated or expanded or when an ancestral population split into two paternal populations and migrated into different areas.

Y-chromosome-based phylogeny was used to explore the population origin, admixture, and evolutionary history at the era of the first genome sequencing era. Zerjal et al. genotyped over 32 markers among 2123 males and identified the genetic legacy of the Mongol western expansion [12]. Similar studies focused on large-scale population cohorts in Mongolia and China revealed that regional population migration and admixture models contributed to the observed patterns of genetic diversity rather than simple cultural diffusion [13, 14]. The significant advance changes in the research patterns occurred when Wei et al. introduced the high-coverage complete Y-sequences to identify the novel phylogenetic variations and calibrate the Y-chromosomal phylogeny [15]. They reported 6662 high-confidence lineage informative SNPs (LISNPs) by analyzing 8.97 Mb of the NRY regions in 36 diverse genomes. Followingly, Karmin et al. sequenced 456 geographically diverse individuals and found that global cultural changes were associated with the identified bottleneck of Y-chromosome diversity based on the phylogenetic analysis of the high-coverage Y-chromosome sequences [16]. Poznik et al. focused on the 1244 whole Y-chromosome genomes from the 1000 Genomes Project and observed punctuated bursts in human male demography inferred from the identified ~ 6000 variants [7]. Large-scale Y-chromosome phylogeny reconstruction from the European, Oceanian and Siberian populations further reported complex population origin tracts, gene flow events, population standstill, rapid population divergence and expansion [17–20]. The single population-scale Y-chromosome investigations focused on Chinese Tibetan, Han, Mongolian and other populations have been conducted, which have identified the population-specific founding lineages and corresponding population origin, separation, and following admixture events [10, 21–23]. However, large-scale Y-chromosome surveys based on high-resolution systems or whole-genome sequencing remained to be conducted to present the entire landscape of genetic diversity and fine-scale paternal genetic history.

The deeper understating of Y-chromosome variations in human evolutionary research further promoted its wide applications in forensic science [4, 24]. The Y-chromosomal phylogeny has become an essential pillar of forensic pedigree searches and paternal ancestry inference. On this basis, several Y-SNP panels of different resolutions have been developed and validated in geographically and linguistically diverse populations [25–31]. Due to the popularity of the capillary electrophoresis (CE) platform, currently available dedicated Y-SNP genotyping tools developed based on this system have restrictions on the number of Y-SNPs analyzed simultaneously [26, 28, 29, 32, 33], hindering the dissection of paternal biogeographical ancestry at higher resolution. Here, targeted NGS technologies are up-and-coming, which have the capability to sequence multiple targets and samples simultaneously and can take advantage of a large number of Y-SNPs for forensic investigations on a detailed level. An early proof-of-principle study showed that 530 Y-SNPs could be genotyped simultaneously in a single sequencing run [25]. This Y-SNP NGS panel covered branches of the entire phylogenetic tree (Y-DNA Haplogroup Tree 2013) and could be applied to comprehensive paternal lineage classification. Subsequently, Ralf et al. presented a vastly improved Y-SNP NGS panel covering 859 Y-SNPs and 640 corresponding paternal lineages [34]. Although all major Y-chromosomal lineages were included in these two panels, finer-scale paternal lineages were lacking. Chinese populations with a large population size possessed a large number of terminal paternal lineages, which limited the application of the above two NGS panels in pedigree search, personal identification and biogeographical ancestry inference. To promote the application of Y-SNPs for forensic investigations in China, multiple Y-SNP NGS panels aiming at the paternal lineage classification of ethnolinguistically diverse populations were successively developed [27, 31, 35, 36]. However, Chinese populations possess complex patterns of cultural, geographical, ethnic and genetic diversity. These dedicated Y-SNP genotyping tools could not simultaneously cover the dominant paternal lineages of Chinese populations and meet the requirement of high-resolution paternal lineage classification. Whole-genome sequencing-based genetic studies have illuminated that Chinese population structures were strongly correlated with geographical regions or language families [37, 38]. Ancient DNA of East Eurasians further identified multiple ancestral sources and complex demographic events that contributed to the gene pool of modern Chinese people, including the westward migrations of the steppe pastoralists and herders, north-to-south di-directional population movements along the Yangtze River and Yellow River basins, peopling of the Tibetan Plateau, and extensive interaction with ancient Siberians [39–43]. These ancient population events further complicated the patterns of the Y-chromosomal lineages in China. To explore the fine-scale paternal genetic structure and illuminate the patterns of genetic diversity of Chinese populations, we developed one high-resolution revised Y-SNP panel with high coverage of geographic and ethnic specificity and the high resolution of terminal haplogroup dissection. We genotyped 639 Y-SNPs in 1033 unrelated individuals from 33 Sinitic-speaking Han and Hui, Mongolic-speaking Mongolian, Tai-Kadai-speaking Gelao and Li, and Tungusic-speaking Manchu populations (Fig. 1A). We aggregated our data with previously publicly available data and comprehensively characterized the genetic diversity and population genetic features based on the sequence variations. We constructed one comprehensive revised forensic phylogenetic tree to present the patterns of Chinese Y-chromosome diversity and illuminate the high performance of our developed panel for forensic applications. Our work has conducted one of the most extensive genetic studies to present one high-density Y-chromosomal phylogeny among linguistically representative Chinese populations. We identified extensive Chinese genetic diversity of ethnolinguistically diverse populations, which can be used as a repertoire of Chinese genomic variations for a better understanding of population admixture events in future studies focused on migration, ancestral source, evolution, demography, adaptation and human genomic resources in China.

Results

Genotyping and genetic diversity among Chinese people

We generated genotype data of 639 Y-SNPs from 322 unrelated individuals from six Chinese populations belonging to three different language families using our developed 639-plex Y-SNP panel (Method), including three Mongolic-speaking Mongolian populations in North China and one Sinitic-speaking Hui in Northwest China, two Tai-Kadai-speaking Gelao in Southwest China and Li in southernmost China (Fig. 1A). To provide one comprehensive population reference data from geographically diverse Chinese populations, we genotyped approximately 30 K Y-SNPs in 711 Han, Manchu, Mongolian and Hui people from 27 populations

(See figure on next page.)

Fig. 1 The geographical positions of 33 studied ethnolinguistically diverse populations and genetic features inferred from the haplogroup frequency spectrum. **A** Geographical locations and the haplogroup composition of 32 predefined 4-level haplogroups. All used haplogroups were manually cut at the fourth level to achieve a statistical possibility. **B** The heatmap showed the clustering patterns of 32 cut-haplogroups from 33 populations. **C**, **D** Principal component analysis (PCA) showed the genetic similarities and differences among 33 populations based on the top three components extracted from the haplogroup frequency spectrum (HFS). **E** Multidimensional scaling plots showed the genetic clustering patterns based on the pairwise Fst matrix. **F** Heatmap showed the pairwise Fst calculated based on the HFS and the population clustered patterns **G**, **H** Neighboring-Joining phylogenetic tree reconstructed based on the genetic distances in the different levels of HFS

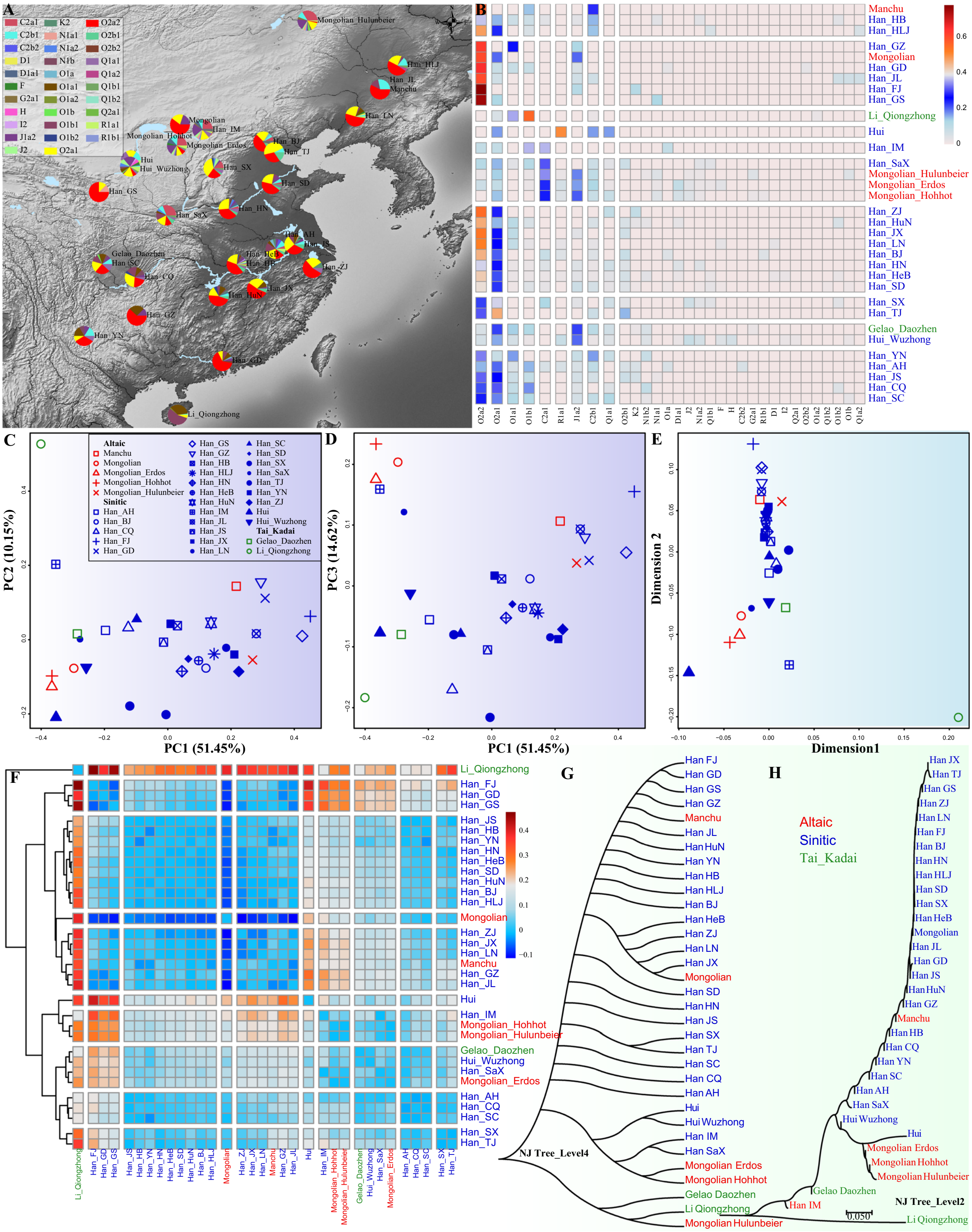

Fig. 1 (See legend on previous page.)

using a high-density array, which included our genotyped 639 Y-SNPs. The average sequencing depth was high enough for our quality control (Method). Among the final dataset of 1033 individuals from 33 populations, we observed 256 definitive Y-chromosomal lineages with the haplogroup frequency ranging from 0.0010 to 0.0687. Eighty-seven haplotypes were observed only once among our dataset, mainly including lineages from Q, R, C, D and other rare O terminal lineages. If we defined a haplogroup frequency larger than 5% as the threshold of the common lineages, the threshold of 1% as low-frequency lineages and other non-singleton as the rare lineages, we observed one common lineage (O2a2b1a1a1a1a1a1a1-M6539, 6.87%, 71/1033), 18 low-frequency lineages (O2a1b1a1a1a1a1a1-F17, O2a2b1a1a1a1a1b1a1b-MF15397, O2a2b2a1b1-A16609, O1b1a1a1a1b2a1a1-F2517, O2a2b1a1a1a1a1a1-F155, O1a1a1b-Z23420, C2a1a2a2b-F12502, O2a1b1a1a1a1-F325, O2a1b1a2a1a1-F1894, O2a2b1a2a1a1a2a1-F273, O2a1b1a1a1a1a1-F110, C2a1a3a-F3796, C2a1a1b1a-F3830, O2a1b1a1a1a1e1a-Y16154, J2-PF4922, Q1a1a1-F1626, O1a1a2a1a-Z23266 and O2b1a1a1-Y173834, the observed number over 11) and 150 rare lineages. We also statistically estimated the distribution of upstream Y-chromosomal lineages. We randomly cut the fourth level of each included haplogroup and observed six common lineages, including C2b1-F2613, O1a1a-CTS4351, C2a1-F3914, O1b1-F2320, O2a1-CTS7638 and O2a2-P201, seven low-frequency lineages, 15 rare lineages and six singletons. Finally, we calculated the nucleotide diversity (π) among 322 individuals from six populations and obtained an estimate of 0.033. The estimated segregating sites of these included Y-SNPs were 256 (the number of sites). The number of sites of parsimony-informative sites was 193. Tajima's D statistic was -1.85776 with a p-value of 0.9856. Analysis of genetic diversity showed our developed panel was suitable for capturing the Chinese population's genetic variations.

Haplogroup frequency spectrum among 33 Chinese populations suggested their differentiated paternal genetic structure

We explored the similarities and differences based on the haplogroup frequency spectrum (HFS) among 33 investigated populations at level 4 (Fig. 1A, B). O1a only existed in Han Chinese, and the common lineage of O1a1 (0.3559) and low-frequency lineage of O1a2 (0.0169) were observed in Tai-Kadai-speaking Li with the highest proportion. Interestingly, Li-dominant lineage O1a1 was commonly identified in southern Han Chinese, Gelao and Hui people. O1b1, with the highest frequency of 0.5593, was mainly observed in southern Han Chinese populations and also identified in northern Han and Mongolian people. Two of four O2 lineages (O2a1 and O2a2) were frequently observed in Han Chinese. O2a1 had the highest frequency in Tianjin Han, but O2a2 had the highest frequency in southern Hans (Fujian, Guangdong and Guizhou). C2a1 widely existed in Mongolian and other northern Han Chinese, and C2b1 was frequently observed in Manchu, Hui and other Han people. Except for the lineages mentioned above, we also observed that Siberian lineages of Q1a1, and western Eurasian-dominant lineages of J1a2, R1a1 and R1b1 contributed to the paternal genetic diversity of Chinese populations (Additional file 1: Tables S1, S2 and Fig. 1A). Heatmap of the HFS showed aforementioned common lineages contributed to the significant components of our studied population's gene pool. We also found that geographically or ethnolinguistically close people shared similar patterns of HFS.

We additionally explored the genetic similarities and differences among 33 Chinese populations using principal component analysis (PCA) based on the HFS on the fourth level. PC1, extracting 51.45% variance from total variations, separated the northern and southern Chinese populations and PC2, extracting 10.15% variance, separated Tai-Kadai-speaking Li from other Chinese people (Fig. 1C). Clustering patterns based on the first and third components separated Mongolian populations from others. Interestingly, Han Chinese populations from Inner Mongolia and Liaoning clustered closely with Mongolian people, suggesting their admixture and extensive interaction status, consistent with the MDS-based (multidimensional scaling analysis) clustering patterns (Fig. 1D, E). Mongolian and Manchu people from the metropolitan populations clustered with Han Chinese, which suggested that southern Altaic people mixed with Han Chinese and other indigenous populations in the historical periods. These observed patterns were consistent with the admixture patterns inferred from our recent genomic studies of Mongolian and Manchu people from Guizhou province [44]. We further validated the identified patterns based on the pairwise Fst matrix, in which Li separated from others, and Mongolian and northern Han clustered together and separated from others. Indeed, the estimated Fst values showed that Li people had the most considerable genetic distances with other comparative populations (Additional file 1: Table S3) and separated from other populations, and formed one isolated clade in the heatmap clustering pattern (Fig. 1F). Other populations were distinguished into two groups, mainly from northern and southern China. Finally, to confirm the robustness of our reconstructed genetic affinity, we reconstructed two Neighboring-Joining trees based on the Fst matrixes at the fourth and second levels (Fig. 1G, H). We found that the phylogenetic relationship inferred from the upstream Y-chromosomal lineages was more consistent with the cluster patterns observed in the PCA, MDS and heatmap.

High-resolution Y-chromosomal lineages for ethnolinguistically diverse Chinese populations

Recent whole-genome sequencing studies of Chinese populations have identified the fine-scale paternal genetic structure of ethnolinguistically different Chinese people and unreported LISNPs [10, 22, 23, 45]. However, whole-genome sequencing for every forensic case sample is impossible. Thus, our developed high-resolution Y-SNP panel is the best choice for promoting forensic applications. We chose all essential lineages in Chinese populations with the divergence times before 500 years. To validate the lineage coverage of our panel, we first genotyped 322 unrelated samples from six populations (Mongolian, Hui, Gelao and Li, Fig. 2A). the reconstructed revised phylogeny among six ethnic groups showed that paternal lineages fell into O2a2, O2a1, O1b1, C2a1, O1a1, C2b1, Q1a1, R1a1 and D1a1, respectively, sampled from 61, 46, 43, 42, 32, 24, 11, 10 and 9 individuals (Fig. 2A). Dominant sublineages (O1b1a1a1a1b2a1a1, 23; O1a1a1b, 12 and C2a1a3a, 11) were observed and restricted to Li and Mongolian populations. Phylogeny results suggested that the C2a/2b can be identified as the founding lineage and ethnicity-specific lineage informative Y-SNP markers of Mongolians for further population genetics and forensic pedigree search as well as biogeographical ancestry inference. Similarly, the identified common sublineages of O1a/1b can be used as the Li-specific founding lineage for subsequent forensic application. We also found some rare lineages originated from Siberia or western Eurasia and participated in the formation of Mongolian and Hui people in North China, suggesting the extensive population admixture along the population migration between North China, Siberia and Central Asia along the silk road or ancient Trans-Eurasian cultural and population communication. Based on the phylogenetic topology, we found that these founding lineages experienced population expansion and the admixture-introduced lineages remained a limited population size in Chinese populations.

To directly explore the genetic connection among different Y-chromosomal lineages and ethnically different populations, we constructed the Network relationship among 322 Chinese males. Consistent with the identified common, rare, or low-frequency lineages in the phylogenetic topology, we observed eleven different lineages contributed to the genetic diversity of these studied populations. C2a1 lineage was unique in Mongolian, and C2b1 was dominant in Mongolian and presented some sublineages in Gelao. O1a/b was dominant in southern Chinese populations (Gelao and Li). O2a2a1/b1, O2a1a1, and O2a1a1/b1 contributed to Mongolian, Hui and Gelao. We could also identify some star-like expansion of some dominant lineages, such as C2a2b1a1a1 in Mongolian populations.

Extensive population expansion and admixture within and between Han Chinese populations and other minorities

Han Chinese is the largest ethnic group in the world and has a significant influence on the formation of the gene pool of Chinese populations [37, 38, 46–48]. To further explore phylogenetic topology among Han Chinese and illuminate the genetic relationship between Han Chinese and aforesaid minorities, we collected and genotyped 711 individuals via Affymetrix array, which included 690 Han Chinese from 25 geographically different populations. We reconstructed a more representative phylogeny using 1033 sequences and confirmed the Mongolian and Li-specific founding lineages (Fig. 3). We found that these two founding lineages participated in the formation of Han Chinese populations. Among Han Chinese populations, we also identified the western Eurasian-introduced R1a/1b and Siberian-mediated Q1. We identified Han Chinese dominant Y-chromosomal lineages of O2a2, O2a1, O1b1, O1a1, C2b1, C2a1, O2b1 and Q1a1, respectively, sampled from 303, 147, 38, 35, 33, 26, 26 and 18 individuals. Some sublineages were also underwent recent population expansion (O2a2b1a1a1a1a1a1, 69; O2a1b1a1a1a1a1, 38; O2a2b1a1a1a1b1a1b, 5; O2a2b1a1a1a1a1, 22; O2a2b2a1b1, 22). The observed mosaic pattern of the identified paternal lineages showed extensive gene flow among Han Chinese populations and minority groups, and the gene flow influence was bi-directional. Mongolian and Li dominant founding lineages were observed in Han Chinese individuals, suggesting that ancient Baiyue ancestors and Eurasian pastoralist people participated in the formation of Han Chinese. Han Chinese dominant lineages were also identified in Mongolian, Hui, Li and Gelao people, which supported population interaction from the paternal perspective.

The reconstructed merged Network confirmed the complex admixture and expansion events among Han Chinese and other non-Han Chinese populations (Fig. 4). The Mongolian and Li-specific lineages (C2a/b and O1a/b) were separated from the Han Chinese-related lineages. We could also find that Li and Mongolian dominant lineages influenced the Han Chinese gene pool as the population composition of one target haplogroup. The distribution of samples belonging to O2a1/a2 denoted that their primary ancestry was derived from Han Chinese populations and minor ancestry from Li. However, the obtained ancestral

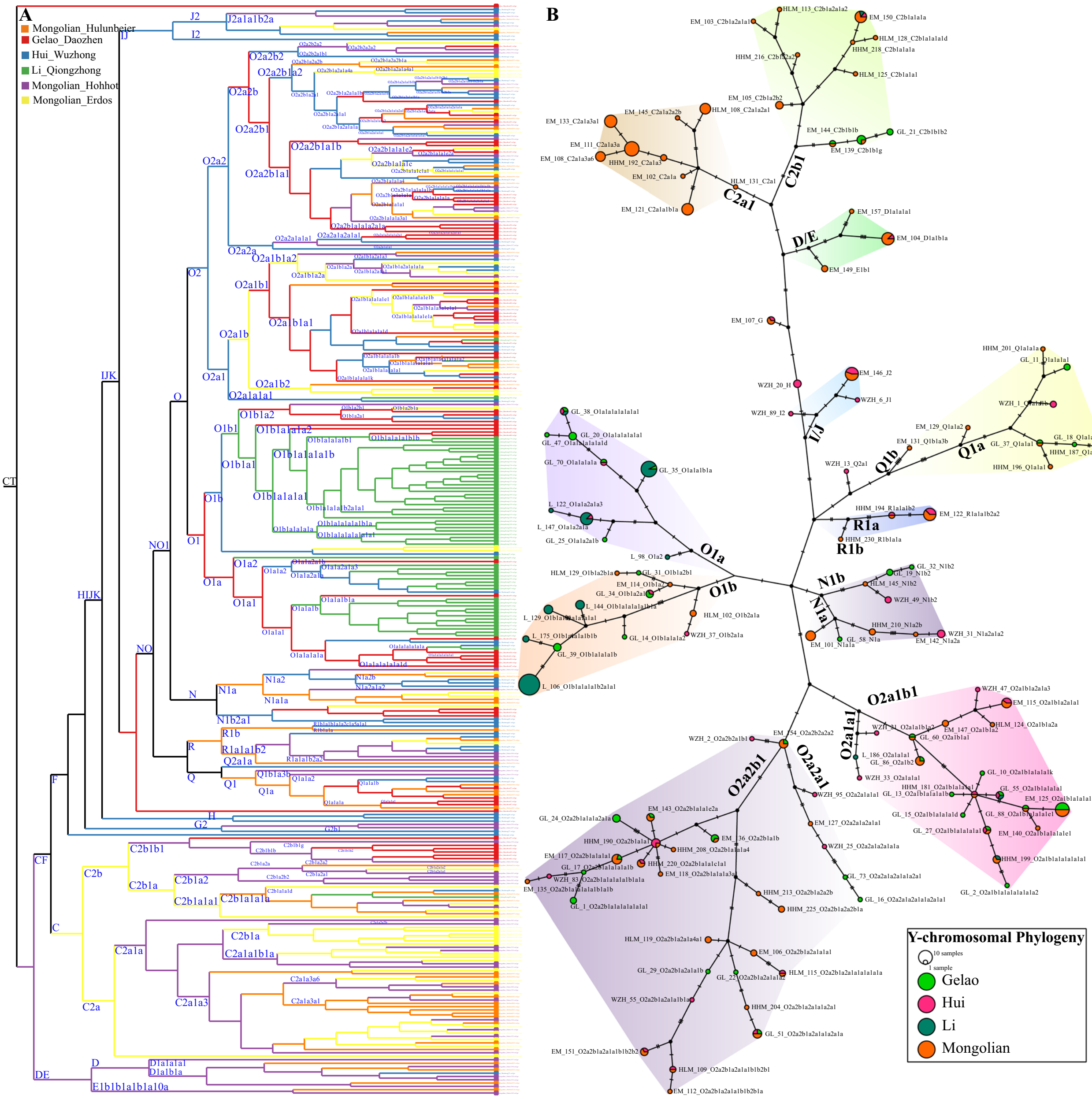

Fig. 2 Phylogeny reconstruction among 322 Chinese individuals from Mongolian, Gelao, Hui and Li people. **A** Phylogenetic relationships were reconstructed based on the Y-LineageTracker. Different colors denoted the diverse populations. **B** The Network reconstructed via the popART showed the shared haplotypes and mutations among other terminal haplogroups. The different colors showed different populations. Different color backgrounds denoted the highlighted Y-chromosomal lineages among Chinese people

lineages from Han Chinese in Gelao, Mongolian, and Hui people were more evident than that in Li people. O2a2b1a1a1a1 and O2a1b1a1a1a1e1 were two important lineages that experienced population expansion.

Paternal genetic connection and differentiation between East Asians and Southeast Asians

Ancient DNA from Island Southeast Asia (Nagsabaran Site) and Mainland Southeast Asia (Man Bac, Nui Nap

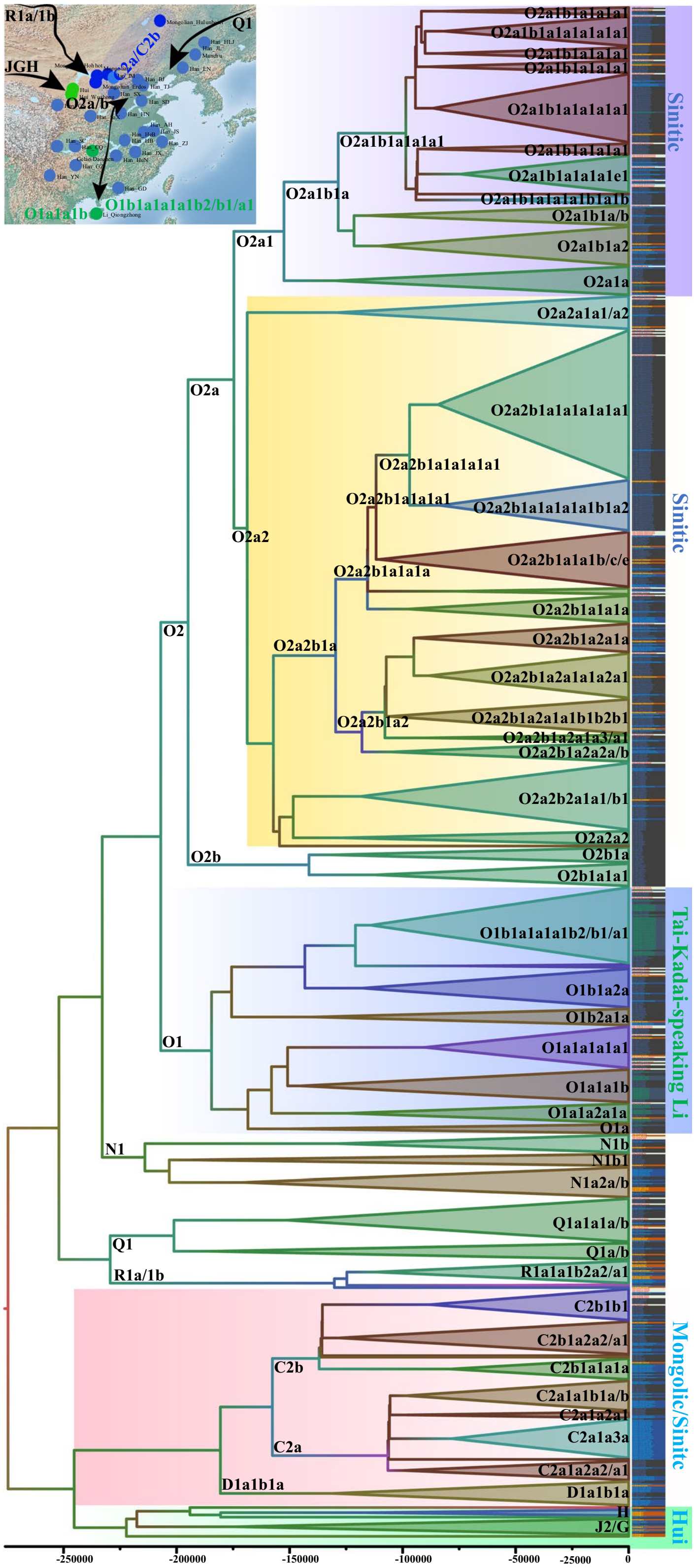

Fig. 3 Phylogenetic relationships based on the sequence diversity among 1033 Chinese males from 33 populations. Different color in a circle showed their different composition of population origins. The different color backgrounds showed different founding lineages. The haplogroup followed by the sample ID was classified using the HaploGrouper

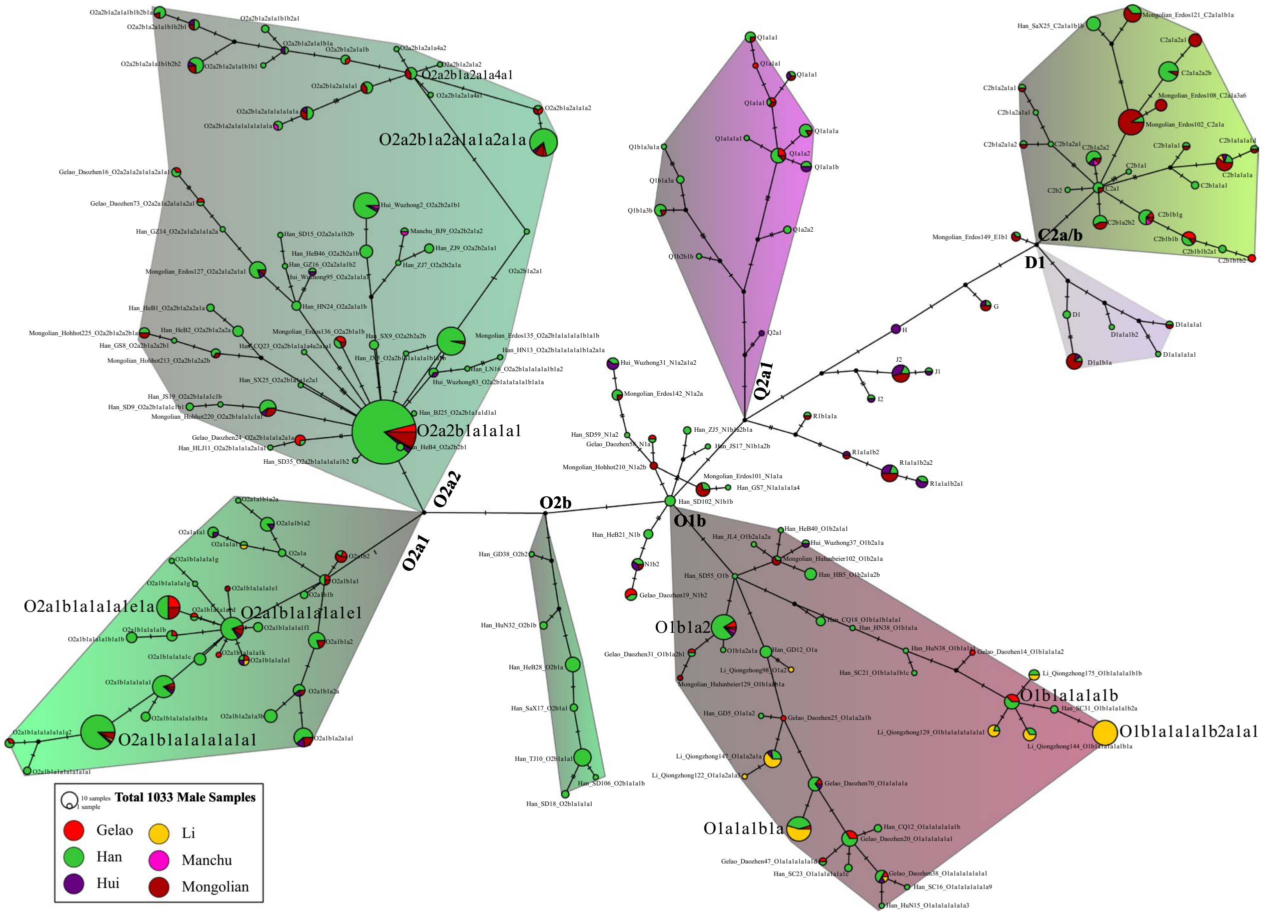

Fig. 4 The reconstructed phylogeny of Chinese populations. The map showed the primary distribution of our investigated Y-chromosomal lineages. The arrow showed the possible migration direction. Circles in the map were coded based on their language families. The line color and triangle denoted the posterior. The triangle length showed the population size and the height showed the relative divergence time. Population ID was categorized based on their ethnicity belongs

and others) has demonstrated that both ancient Southeast Asian Hòabìnhian hunter-gatherers and Neolithic southern Chinese rice farmers contributed to the complexity of genetic diversity of Southeast Asia [49, 50]. Larena et al. recently reported the large-scale genetic diversity of geographically and ethnolinguistically diverse modern Philippine populations and illustrated more complex processes of the peopling history of Island Southeast Asia [51, 52]. Northern and southern Negritos, Manobo, Sama, Papuan, and Cordilleran-related ancestral populations contributed to the gene pool of the modern landscape of Philippines [51], and Ayta people possessed the highest level of Denisovan ancestry among Negritos and Papuan people [52]. Complex population admixture and rich genetic diversity are also reported in Mainland Southeast Asia [53, 54]. Focused on the uniparental history of Southeast Asians, Kutanan and his cooperators genotyped 2.3 MB Y-chromosome variations among ethnolinguistically diverse Southeast Asian and reported their similarities and differences between the paternal and maternal genetic history [11, 55, 56]. However, the genetic interaction between Southeast Asia and East Asia remains to be comprehensively evaluated.

To comprehensively investigate the population interaction between East and Southeast Asians, we aggregated our data with publicly available haplogroup information from 3094 people. We obtained one aggregation dataset that included 4728 individuals from 114 ethnolinguistically populations from China and Southeast Asia, including 5 Altaic-, 6 Austroasiatic-, 5 Austronesian-, 62 Sino-Tibeto-, 10 Hmong-Mien-, 26 Tai-Kadai-speaking populations (Additional file 1: Table S4). We first conducted PCA analysis based on the haplogroup frequency at the second level and found the top three components extracted 77.84% variances from the total variations. Generally, patterns inferred from Y-chromosomal haplogroup variations were more obscure than that inferred based on the whole-genome autosomal variations. Nevertheless, we can still identify that some ethnolinguistically specific people were separated from others (Fig. 5A, B). Apparent population affinity within similar language families can be identified in the pairwise Fst heatmap, where the lowest genetic distances were observed among linguistically close populations (Fig. 5C). We further explored the haplogroup composition of major lineages among 114 populations, and we found that the dominant lineages were enriched in some linguistically specific or geographically isolated populations. F lineages were observed in Lahu and Phula, and D lineages were dominant in Yi and Tibetan. O1 was dominant in Southeast Asians, and O2 was dominant in Chinese populations (Fig. 5D). We interestingly identified that Hmong people from Southeast Asia possessed the largest proportion of O2 lineage, which was consistent with the long-distance population migration and connection inferred from the genome-wide SNP data [57]. The reconstructed phylogenetic relationship based on the pairwise Fst matrix showed two major branches associated with the stratification of haplogroup composition of Southeast and East Asia (Fig. 5E). We also found that Hui people possessed complex haplogroup composition and clustered together with Han Chinese branch. Generally, our comprehensive population comparison showed the close genetic connection between Southeast Asian and southern Chinese populations and differentiation between Southeast Asian and northern Chinese groups, especially with Altaic-speaking populations.

Discussion

The full landscape of Y-chromosomal diversity reveals complex population migration and admixture tracts

Non-crossover regions of the human Y-chromosome harbor the feature of male-specific inheritance and can record most male behavior, phenotype and human demographic details [4, 7]. To explore the patterns of Y-chromosomal diversity, we reported the genotypes of 1033 Y chromosomes randomly sampled from 33 Chinese populations belonging to five ethnic groups (Mongolian, Manchu, Hui, Gelao, Li and Han), which were genotyped using our newly-developed 639-plex Y-SNP panel and high-density Affymetrix array. We have conducted a comprehensive population evolutionary analysis and population comparison tests within and between Chinese populations belonging to different geographical regions or language families. Population genetic survey suggested that our panel captured the richest Y-chromosomal genetic diversity to date in all forensic Y-SNP genotyping tools focused on the Chinese populations [26, 27, 31, 32, 36, 58–60]. Phylogeny constructed among Non-Han Chinese (Mongolic, Hui and Tai-Kadai) and all included subjects consistently demonstrated that strong geography or ethnicity-related Y-chromosomal features indicated the underlying complex population evolutionary history and potential for forensic pedigree search and biogeographical ancestry inference.

HFS analysis among our studied populations or regional northern Mongolian and southern island Li people has revealed their common founding lineages of C2a/2b and O1a/1b (Fig. 1A, B). Geographical distribution further confirmed that these dominant lineages could be used as forensic markers for genetic localization. C2a1-F3914 can be used as a Mongolian predominant founding lineage, which was observed in 42 Mongolian individuals and 26 Han Chinese individuals, mainly from Shaanxi, Shanxi and Inner Mongolia. Another Mongolian predominant lineage C2b1-F2613 was observed in 16 Mongolians (16/159), two Manchu, one Li, two Hui, 32 Han (32/693) and six Gelao individuals. Upstream C2b-F1067 (previously classified as C2c) was reported first in northeastern Asia and associated with the origin and expansion of Mongolic-speaking populations. Subclades of C2b1a1a1a-M407 (C2c1a1a1 in the previous version) appeared in ten individuals (two Hans, one Hui, one Li and four Mongolians), and C2b1b1b-F5477 and C2b1a2b2-FGC45548 were also, respectively, observed seven and six times. Huang et al. presented one revised phylogenetic tree and distribution map focused on all available C2b1a1a1a-M407 samples and found that C2b1a1a1a-M407 has a frequency of over 50% in the northeastern Asian populations [61]. Thus, C2b1-F2613 and C2b1a1a1a-M407 can be used to trace the population origin and migration of Kalmyks, Mongolians, Buryats and other genetically close northeastern Asians.

Network and phylogeny constructions further showed that the lineage of O1a-M119 was the common paternal lineage in southern Chinese populations (Fig. 2). Sublineages of O1b1a1a1a1b2a1a1-F2517 (23), O1b1a2a1-F1759 (10), O1a1a1b-Z23420 (15) and O1a1a2a1a-Z23266 (11) had undergone population expansion recently. All F2517 lineages were observed in Li people, which was consistent with a recent whole-genome sequencing study [62]. Chen et al. found O1b1a1a was the dominant

(See figure on next page.)

Fig. 5 Genetic connection and differentiation between Southeast Asians and East Asians inferred from Y-chromosome variations. **A**, **B** PCA inferred from the top three components showed the genetic connections among 114 populations. **C** Heatmap of pairwise Fst genetic distances among 114 populations. **D** Haplogroup composition of major Y-chromosomal lineage from 4727 individuals from 114 populations. **E** The phylogenetic relationship among 114 populations showed their genetic affinity. The branch length was not associated with the genetic drift. All populations from one family, ethnicity or geographically close regions were color-coded via one color

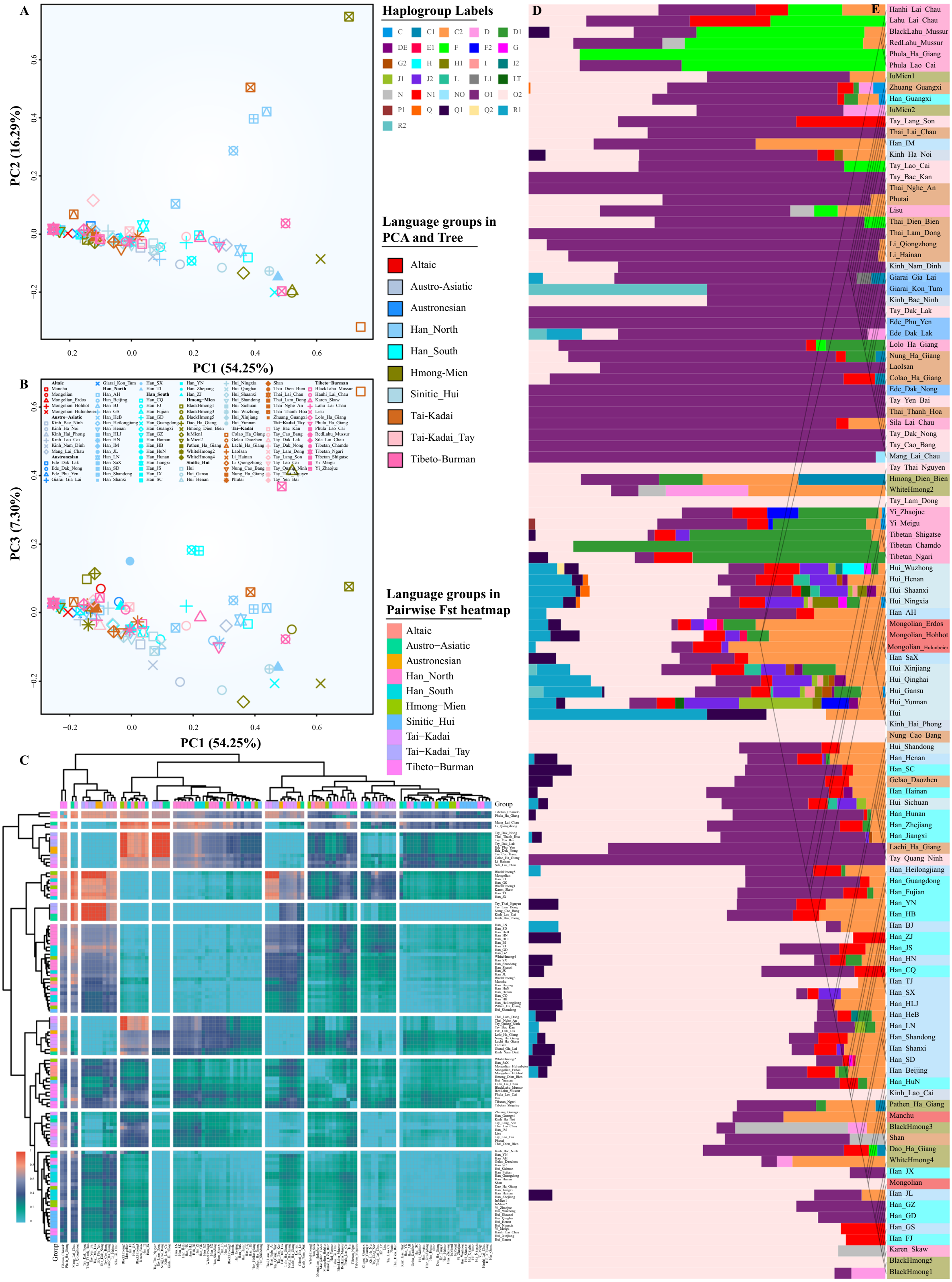

Fig. 5 (See legend on previous page.)

lineage in southern East Asians, which diverged from others 10,998 years ago, and the F2517 sublineages were further divided into O1b1a1a1a1b2a1a1a and O1b1a1a1a1b2a1a1b clades at 2828 years ago [62]. Sun et al. also found that most sublineages of M119, including our identified F1759, Z23420 and Z23266, contributed to the ancient gene pool of modern Tai-Kadai, Austronesian and southern Han Chinese [10]. We could also identify the shared paternal lineages among Li, southern Han and Gelao people in the distribution of M119 mutations. The unique paternal genetic structure of the Hainan Li people was also consistent with our previous findings of the fine-scale genetic structure [63].

Our survey has identified 195 samples of O2a1-CTS7638 and 372 samples of O2a2-P201 in Han Chinese populations. Sublineages of O2a2b1a1a1a1a1a1-M6539 (71), O2a1b1a1a1a1a1-F17 (43), O2a2b1a1a1a1b1a1b-MF15397 (26), O2a2b2a1b1-A16609 (24) and O2a2b1a1a1a1a1-F155 (23) have experienced population expansion events recently, which was consistent with population expansion among Han Chinese populations inferred from the whole-genome sequencing project. Admixture-introduced rare lineage Q1a1-F746/F790 was observed in 30 samples from Mongolian, Han, and Gelao people, and steppe pastoralist-related R1a1a1b2-Z93 was observed in 17 Mongolian and Hui people. Our results provided genetic evidence for extensive admixture between northern East Asians and surrounding populations. Similar patterns were also illuminated in the paternal genetic history reconstruction by Wang et al., who concluded that multiple ancestral sources contributed to the formation of the paternal gene pool of Mongolian people [32]. Besides, He et al. found that Mongolic-speaking populations have strong population stratifications, in which the northern one was influenced by Siberian ancestry, the western one was influenced by western Eurasian and the southern one was influenced by Han Chinese expansion [37]. Recent ancient DNA also found that western Eurasian steppe ancestry has influenced the genetic makeup of northern East Asians. For western Eurasian ancestry identified in Hui people, our paternal results were also consistent with the admixture patterns via the genome-wide SNP data. Complex demographical models suggested that geographically diverse Chinese Hui people harbored complex and different admixture processes and possessed approximately 10% ancestry related to the ancestor from Central Asians [64–66].

We also identified paternal genetic structure among Chinese populations in the clustering patterns via HFS, PCA and MDS. These population stratifications were in accordance with the language or geographical categories. Island Li people shared their specific paternal genetic structure and clustered far from other Chinese populations. Similarly, Mongolian people clustered together and have a relatively close genetic relationship with northern East Asians. We also should note that the Y-chromosome-based population structure in China is rougher than that inferred from the genome-wide SNPs. Our recent genetic studies have identified five population substructures correlated with languages and geography in China. Mongolic and Tungusic people in northeastern China harbored the highest Ulchi or ancient Boisman/DevilsCave-related ancestry [37, 41]. Tibeto-Burman groups from Tibetan Plateau had the highest proportion of ancestry related to core Tibet Tibetan and Nepal Chokhopani, Mebrak, and Samdzong people [67, 68]. The primary ancestral component of Hmong-Mien people from southwestern China was maximized in ancestry related to Miao and Yao people [69], and Austronesian-speaking people from Taiwan Island have more Ami/Atayal or ancient Hanben-related ancestry [41]. Han Chinese ancestry localized between the four ancestries mentioned above and showed a northern-to-southern genetic cline. The recent large-scale genetic structure also identified fine-scale population structure among geographically diverse Han Chinese populations [46, 47, 70]. Our population comparison among Southeast Asians and East Asians also identified the shared genetic material, such as O1 in ingenious southern people and O2 in Hmong-Mien people. These patterns were consistent with multiple waves of migrations from South China to Southeast Asia inferred from ancient DNA [49, 50] and previous modern DNA from Southeast Asia [11, 51, 53–56]. These fine-scale genetic backgrounds could promote better study design for large medical clinical cohorts and forensic genetic localization of crime cases.

639-plex Y-SNP panel can be used as a powerful forensic tool for Chinese forensic pedigree search and biogeographical ancestry inference

The forensic community has noticed that whole-genome sequencing in a forensic case needs to overcome specific infrastructure of the specialist, platform and genomic statistical methods, as well as the experiment method focused on the forensic case samples. Besides, the cost of one sample is another important obstacle to the wide application of whole-genome sequencing technology in forensics. Evolutionary genetic scientists have conducted many vital projects to explore the complete anthropologically-informed phylogeny [7, 19, 20]. Our work has identified most paternal founding lineages in Chinese populations and comprehensively characterized their geographical and ethnic distribution. Our panel harbored the high coverage of genetic variations of terminal lineages and complete coverage of reference data from main populations or ethnic groups in China.

Forensic pedigree search can help trace possible crime suspects based on the shared Y-chromosome mutations. Lineages informative Y-SNPs were usually used together with Y-STR markers. Many prior works provided relatively high-resolution forensic phylogenic trees and presented the corresponding scientific examination and analysis strategies. They promoted the advances of forensic Y-chromosome applications in pedigree search and biogeographical ancestry inference based on the customized SNaPshot and NGS technologies [26, 27, 31, 32, 36, 58–60]. Song et al. explored the paternal genetic structure of Hainan Li using their developed panel containing 141 Y-SNPs. They found that haplogroup O1b1a1a1a1a1b-CTS5854 can be used as one ethnicity-specific lineage in population and forensic genetics [59]. Song et al. further updated their panel, including 233 Y-SNPs used for Chinese Qiang people, and found that O2a2b1a1-M117, O2a2b1a1a1-F42 and O2a1b1a1a1a-F11 were the founding lineages in Qiang people [60]. Wang et al. also investigated the paternal genetic structure of Zhuang people using this panel and identified the O2-dominant lineages in Tai-Kadai people [58]. Xie et al. developed one panel focused on Hui people, which included 157 Y-SNP, and identified the population substructure of Hui people [26]. Wang et al. focused on the genetic diversity of Mongolian people (N1b-F2930, N1a1a1a1a3-B197, Q-M242 and O2a2b1a1a1a4a-CTS4658) and developed one Mongolian-specific panel included 215 Y-SNPs [32]. The panels mentioned above consisted of several customized SNaPshot systems, which limited the rapid use in forensic cases. Wang et al. developed one 165-plex Y-SNP panel based on an Ion S5 XL system. They comprehensively conducted the sequencing performance and concordance, reliability, sensitivity, and stability studies based on the ISFG guidelines [27]. Liu et al. recently updated this system by increasing the final Y-SNP number to 256 [36]. Significantly, Tao et al. developed a customized SifaMPS Y-SNP panel that included 381 Y-SNPs focused on Chinese populations and investigated the basic structure and sub-branches of Chinese major haplogroup branches [31]. Our panel included two significant features: the first one is that lineages specific to or common in most Chinese populations were included (O, D, C, R and Q et al.), and the other important one is that we retained a higher resolution of the terminal Y-chromosomal lineages, which can complete the shortcoming of previously developed panels limited to some common lineages or only focused on specific populations. The newly-developed panel overcame the limitation of the lineage's representatives, terminal lineage resolution and sequencing platforms, which can provide the best practice tool in forensic applications. The identified paternal population structure in China can provide more clues for biogeographical ancestry inference, which can be used as a complementary tool for forensic ancestry prediction based on the autosome-based ancestry informative SNP panel [64].

Conclusion

The complete landscape of human Y-chromosome variations and the gradually updated Y-chromosome phylogenetic tree with more population-specific LISNPs formed the fundamental for forensic application and evolutionary study. To overcome the shortness of the whole Y-chromosome sequencing in forensic science at the initial stage of the genome sequencing era, we developed one high-resolution 639-plex Y-SNP panel that included 639 LISNPs defined 573 terminal Y-chromosomal lineages. We generated the population data from 1033 individuals from 33 populations, including Han, Hui, Mongolian, Li and Gelao and investigated the forensic features, HFS and evolutionary processes via multiple statistical models. We identified 257 terminal Y-chromosomal lineages with several common founding lineages of O2a2b1a1a1a1a1a1-M6539, O2a1b1a1a1a1a1-F17, O2a2b1a1a1a1b1a1b-MF15397, O2a2b2a1b1-A16609, O1b1a1a1a1b2a1a1-F2517 and O2a2b1a1a1a1a1a1-F155. Patterns of HFS and corresponding geographical distribution illuminated that some Siberian or western Eurasian-originated paternal lineages contributed to the formation of the paternal gene pool of Mongolian, Hui and northern Han Chinese populations. Network and our reconstructed forensic phylogenic topology further illuminated the complex population divergence and expansion of different paternal lineages, which also found some ancestral lineages shared by geographically or linguistically different Chinese populations. PCA and MDS clustering patterns showed that the paternal genetic structure was correlated with the geographical and linguistic categories, which provided the basic genetic background for forensic paternal biogeographical ancestry inference. Our reconstructed revised phylogeny and comprehensive population genetic investigation based on this Y-SNP panel can provide the highest resolution of the terminal lineage and genetic diversity, which provides one panel with high marker coverage and lineage representation. Ethnolinguistically diverse Chinese populations had the highest genetic diversity. Thus, anthropologically-informed Y-chromosome whole-genome sequencing will promote the further development of higher-resolution Y-SNP NGS panels and corresponding population-specific dataset construction. We also emphasized that the large-scale population cohorts, such as 10,000 Chinese Person Genomic Diversity Project (10K_CPGDP) and 100 K-GSRD^{WCH} (100 K genome sequencing of rare disease), can provide more unreported LISNPs for forensic application of Y-chromosome.

Materials and methods

Sample collection and DNA extraction

We collected peripheral blood samples from 1033 unrelated Han, Tibetan, Hui, Gelao, Manchu and Li individuals from 33 geographically different regions (Fig. 1A). Each donor provided written informed consent. The medical ethics boards at Sichuan University and North Sichuan Medical College have approved our study protocol. Our experiments followed the recommendations and regulations of our institute and national guidelines of standards of the Declaration of Helsinki [71]. PureLink Genomic DNA Mini Kit (Thermo Fisher Scientific, Waltham, USA) was used to extract the genomic DNA. Based on the official manufacturer's guidelines, Quantifiler Human DNA Quantification Kit (Thermo Fisher Scientific) and 7500 Real-time PCR System (Thermo Fisher Scientific) were used to quantify the DNA quantity and finally reserved in a low-temperature environment.

Marker composition and NGS-based panel development

We chose the final included markers based on the following five rules to present a full-scale Chinese Y-chromosome diversity and high resolution of each terminal lineage. First, all major clade lineages recorded in the International Society of Genetic Genealogy (ISOGG) Y-DNA Haplogroup Tree 2019–2020 version 15.73 (https://isogg.org/tree/index.html) and Yfull databases focused on Chinese populations were included. Second, key mutations included in our previously developed panel and validated in the population genetic studies were included [27, 36]. Third, we determined the terminal mutations based on the population-scale HFS according to the revised phylogenetic tree in the whole-genome sequencing projects. Based on the public data from the expanded 1000 Genomes Project cohort [3], Human Genetic Diversity Project (HGDP) [2], Simons Genome Diversity Project [72], Estonian Biocentre Human Genome Diversity Panel (EGDP) [73] and 10K_CPGDP and others, we have built one in-house Y-chromosome population database with detailed HFS distribution and estimated divergence times (will be published soon). We choose the terminal mutations with a frequency larger than 5%. Fourth, based on the representatives of the included individuals and haplogroup coverage, we estimated the divergence times of each branch based on the localization of the mutations in the revised phylogeny. We included the mutations with a divergence time older than 500 years. We included 1000 Y-SNPs for the final primer design. We finally developed the Y-SNP NGS panel based on the MGISEQ-2000RS (MGI Tech Co., Ltd., Shenzhen, Guangdong, China) sequencing platform, which has been formally validated based on the SWGDAM guidelines (paper in preparation).

Genotyping and quality control

We sequenced 639 Y-SNPs in 322 samples from Mongolian populations in three geographically different regions (154 males), Hui people from Wuzhong (50 males), Gelao people from Daozhen (59 males) and Li people (59 males) from Hainan, using our developed 639-plex Y-SNP panel on the MGISEQ-2000RS sequencing platform. In each sequencing run, positive control of 2800M (Promega, Madison, WI, USA) and negative ddH$_2$O control were used. To provide one comprehensive comparative database, we also genotyped 639 Y-SNPs in 711 Han, Mongolian and Manchu individuals using an Affymetrix array. We used the PLINK v1.90b6.26 64-bit (2 Apr 2022) to conduct quality control on the merged dataset based on the missing SNP rate and missing genotyping rate with two parameters (−geno: 0.1 and −mind: 0.1) [74]. We finally kept a dataset of 639 Y-SNPs in 1033 unrelated individuals from 33 populations in the following forensic effectiveness evaluation and evolutionary history reconstruction.

Classifying NRY haplogroups

We manually classified the NRY haplogroups for sequencing data based on the predefined phylogenetic tree with the chosen mutation markers. And then, we merged the sequencing data and the chip-based data and used the python package of hGrpr2.py instrumented in HaploGrouper to classify the haplogroups [75]. All branches in the haplogroup tree (treeFileNEW_isogg2019.txt) and ISOGG SNP file (snpFile_b38_isogg2019.txt) were used in the HaploGrouper-based haplogroup classification. We additionally used the chip version (−chip) in the LineageTracker to classify the NRY haplogroups based on the GRCH38 reference genome [76].

Haplogroup frequency spectrum estimation and clustering analysis

We calculated the HFS at different levels of the terminal haplogroups. We estimated the geographical distribution via a pie chart in the map and a heatmap based on the HFS matrix. Followingly, we conducted PCA based on the HFS matrix. We used the top three components extracted from the total variations to cluster our studied populations. We also calculated the pairwise Fst genetic matrix based on the HFS and conducted the

MDS to explore the genetic affinity between included populations.

Phylogeny analysis for NRY haplogroups

We used the LineageTracker to construct the phylogenetic topology of all individuals [76]. Tools for variant calling and manipulating VCFs and BCFs (Bcftools Version: 1.8) and PLINK v1.9 were used to convert the vcf files to fasta files. We used BEAUti to convert fasta files into XML files and ran the BEAST analysis using BEAST2.0 [77]. Tracer v1.7.2 was used to evaluate the power of statistical parameters and TreeAnnotator was used to choose the best trees in the BEAST results [77]. We finally used FigTree v1.4.4 to visualize and organize the phylogenetic tree [77].

Network analysis for Y-SNP haplotype data

Python package of fasta_to_nexus_Main.py (https://github.com/rubenAlbuquerque/fasta_nexus_converter/blob/master/fasta_to_nexus/Main.py) was used to generate the nexus files. We used popART to construct the Network relationship. Here, five Network models were used to build the phylogenetic relationship among different lineages, including Minimum spanning network, Median Joining Network, Integer NJ Net, Tight Span Walker and TCS Network [78]. AMOVA was conducted to explore the genetic similarities and differences between or within groups and populations. Nucleotide diversity was also estimated using the popART [78].

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s40246-023-00476-6.

Additional file 1: Table S1 Haplogroup distribution of all studied populations. **Table S2** Haplogroup frequency of 33 Chinese populations. **Table S3** Genetic distances among 33 Chinese populations. **Table S4** Haplogroups of 4727 people from 114 populations.

Acknowledgments
We thank all volunteers participated to this project.

Author contributions

GLH, MGW, LM, LHW, CL, JY and LW conceived and designed the study. GLH, MGW, LM, JC, JZ, QXS, SHD, ZYW, XFX, YTS, YL, JL, ZW, LHW, CL, JY and LW made all analyses. GLH, MGW, CL, JY and LW wrote the manuscript. GLH and MG revised the paper. All authors read and approved the final manuscript.

Funding
This work is funded by the National Natural Science Foundation of China (82202078) and the Open project of Key Laboratory of Forensic Genetics of Ministry of Public Security (2022FGKFKT05).

Availability of data and materials

We submitted allele frequency data in the Supplementary materials. Following this project's regulations, informed consent, and the Human Genetic Resources Administration of China (HGRAC) regulations, we can share the obtained genotype data via personal communication with the corresponding authors. We make the data available upon request by asking the person requesting the data to agree in writing to the following restrictions: 1, the data can be only used for studying population history; 2, the data cannot be used for commercial purposes; 3, the data cannot be used to identify the sample donors; 4, the data cannot be used for studying natural/cultural selections, medical or other related studies.

Declarations

Ethics approval and consent to participate

Each donor provided written informed consent. The medical ethics boards at Sichuan University and North Sichuan Medical College have approved our study protocol. Our experiments followed the recommendations and regulations of our institute and national guidelines of standards of the Declaration of Helsinki.

Consent for publication

Not applicable.

Competing interests

The author declares no conflict of interest.

Author details
[1]Institute of Rare Diseases, West China Hospital of Sichuan University, Sichuan University, Chengdu 610041, China. [2]Faculty of Forensic Medicine, Zhongshan School of Medicine, Sun Yat-Sen University, Guangzhou 510275, China. [3]National Engineering Laboratory for Forensic Science, Key Laboratory of Forensic Genetics of Ministry of Public Security, Institute of Forensic Science, Ministry of Public Security, Beijing 100038, China. [4]School of Forensic Medicine, Shanxi Medical University, Jinzhong 030001, China. [5]Department of Forensic Medicine, College of Basic Medicine, Chongqing Medical University, Chongqing 400331, China. [6]School of Basic Medical Sciences, North Sichuan Medical College, Nanchong 637000, China. [7]School of Forensic Medicine, Kunming Medical University, Kunming 650500, China. [8]Institute of Forensic Medicine, West China School of Basic Science and Forensic Medicine, Sichuan University, Chengdu 610041, China. [9]School of Ethnology and Anthropology, Inner Mongolia Normal University, Hohhot 010028, Inner Mongolia, China. [10]Guangzhou Key Laboratory of Forensic Multi-Omics for Precision Identification, School of Forensic Medicine, Southern Medical University, Guangzhou 510515, China.

Received: 12 January 2023 Accepted: 15 March 2023
Published online: 28 March 2023

References

1. Almarri MA, Bergstrom A, Prado-Martinez J, Yang F, Fu B, Dunham AS, Chen Y, Hurles ME, Tyler-Smith C, Xue Y. Population structure, stratification, and introgression of human structural variation. Cell. 2020;182(1):189–99.
2. Bergstrom A, McCarthy SA, Hui R, Almarri MA, Ayub Q, Danecek P, Chen Y, Felkel S, Hallast P, Kamm J, et al. Insights into human genetic variation and population history from 929 diverse genomes. Science. 2020;367(6484):eaay5012.
3. Byrska-Bishop M, Evani US, Zhao X, Basile AO, Abel HJ, Regier AA, Corvelo A, Clarke WE, Musunuri R, Nagulapalli K, et al. High-coverage whole-genome sequencing of the expanded 1000 genomes project cohort including 602 trios. Cell. 2022;185(18):3426–40.
4. Jobling MA, Tyler-Smith C. Human Y-chromosome variation in the genome-sequencing era. Nat Rev Genet. 2017;18(8):485–97.
5. Kayser M. Forensic use of Y-chromosome DNA: a general overview. Hum Genet. 2017;136(5):621–35.
6. Yao H, Wen S, Tong X, Zhou B, Du P, Shi M, Jin L, Li H. Y chromosomal clue successfully facilitated the arrest of Baiyin serial killer. Sci Bull. 2016;61(22):1715–7.
7. Poznik GD, Xue Y, Mendez FL, Willems TF, Massaia A, Wilson Sayres MA, Ayub Q, McCarthy SA, Narechania A, Kashin S, et al. Punctuated bursts in human male demography inferred from 1,244 worldwide Y-chromosome sequences. Nat Genet. 2016;48(6):593–9.

8. van Oven M, Van Geystelen A, Kayser M, Decorte R, Larmuseau MH. Seeing the wood for the trees: a minimal reference phylogeny for the human Y chromosome. Hum Mutat. 2014;35(2):187–91.
9. Hallast P, Batini C, Zadik D, Maisano Delser P, Wetton JH, Arroyo-Pardo E, Cavalleri GL, de Knijff P, Destro Bisol G, Dupuy BM, et al. The Y-chromosome tree bursts into leaf: 13,000 high-confidence SNPs covering the majority of known clades. Mol Biol Evol. 2015;32(3):661–73.
10. Sun J, Li YX, Ma PC, Yan S, Cheng HZ, Fan ZQ, Deng XH, Ru K, Wang CC, Chen G, et al. 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. 2021;174(4):686–700.
11. Kutanan W, Kampuansai J, Srikummool M, Brunelli A, Ghirotto S, Arias L, Macholdt E, Hubner A, Schroder R, Stoneking M. Contrasting paternal and maternal genetic histories of thai and lao populations. Mol Biol Evol. 2019;36(7):1490–506.
12. Zerjal T, Xue Y, Bertorelle G, Wells RS, Bao W, Zhu S, Qamar R, Ayub Q, Mohyuddin A, Fu S, et al. The genetic legacy of the Mongols. Am J Hum Genet. 2003;72(3):717–21.
13. Xue Y, Zerjal T, Bao W, Zhu S, Lim SK, Shu Q, Xu J, Du R, Fu S, Li P, et al. Recent spread of a Y-chromosomal lineage in northern China and Mongolia. Am J Hum Genet. 2005;77(6):1112–6.
14. Wen B, Li H, Lu D, Song X, Zhang F, He Y, Li F, Gao Y, Mao X, Zhang L, et al. Genetic evidence supports demic diffusion of Han culture. Nature. 2004;431(7006):302–5.
15. Wei W, Ayub Q, Chen Y, McCarthy S, Hou Y, Carbone I, Xue Y, Tyler-Smith C. A calibrated human Y-chromosomal phylogeny based on resequencing. Genome Res. 2013;23(2):388–95.
16. Karmin M, Saag L, Vicente M, Wilson Sayres MA, Jarve M, Talas UG, Rootsi S, Ilumae AM, Magi R, Mitt M, et al. A recent bottleneck of Y chromosome diversity coincides with a global change in culture. Genome Res. 2015;25(4):459–66.
17. Francalacci P, Morelli L, Angius A, Berutti R, Reinier F, Atzeni R, Pilu R, Busonero F, Maschio A, Zara I, et al. Low-pass DNA sequencing of 1200 Sardinians reconstructs European Y-chromosome phylogeny. Science. 2013;341(6145):565–9.
18. Bergstrom A, Nagle N, Chen Y, McCarthy S, Pollard MO, Ayub Q, Wilcox S, Wilcox L, van Oorschot RA, McAllister P, et al. Deep roots for aboriginal Australian Y chromosomes. Curr Biol. 2016;26(6):809–13.
19. Pinotti T, Bergstrom A, Geppert M, Bawn M, Ohasi D, Shi W, Lacerda DR, Solli A, Norstedt J, Reed K, et al. Y Chromosome sequences reveal a short beringian standstill, rapid expansion, and early population structure of native American founders. Curr Biol. 2019;29(1):149–57.
20. Karmin M, Flores R, Saag L, Hudjashov G, Brucato N, Crenna-Darusallam C, Larena M, Endicott PL, Jakobsson M, Lansing JS, et al. Episodes of diversification and isolation in island Southeast Asian and near Oceanian male lineages. Mol Biol Evol. 2022;39(3):045.
21. Wang LX, Lu Y, Zhang C, Wei LH, Yan S, Huang YZ, Wang CC, Mallick S, Wen SQ, Jin L, et al. Reconstruction of Y-chromosome phylogeny reveals two neolithic expansions of Tibeto-Burman populations. Mol Genet Genomics. 2018;293(5):1293–300.
22. Sun N, Ma PC, Yan S, Wen SQ, Sun C, Du PX, Cheng HZ, Deng XH, Wang CC, Wei LH. Phylogeography of Y-chromosome haplogroup Q1a1a-M120, a paternal lineage connecting populations in Siberia and East Asia. Ann Hum Biol. 2019;46(3):261–6.
23. Liu BL, Ma PC, Wang CZ, Yan S, Yao HB, Li YL, Xie YM, Meng SL, Sun J, Cai YH, et al. Paternal origin of Tungusic-speaking populations: Insights from the updated phylogenetic tree of Y-chromosome haplogroup C2a–M86. Am J Hum Biol. 2021;33(2):e23462.
24. Jobling MA, Tyler-Smith C. The human Y chromosome: an evolutionary marker comes of age. Nat Rev Genet. 2003;4(8):598–612.
25. Ralf A, van Oven M, Zhong K, Kayser M. Simultaneous analysis of hundreds of Y-chromosomal SNPs for high-resolution paternal lineage classification using targeted semiconductor sequencing. Hum Mutat. 2015;36(1):151–9.
26. Xie M, Song F, Li J, Lang M, Luo H, Wang Z, Wu J, Li C, Tian C, Wang W, et al. Genetic substructure and forensic characteristics of Chinese Hui populations using 157 Y-SNPs and 27 Y-STRs. Forensic Sci Int Genet. 2019;41:11–8.
27. Wang M, Wang Z, He G, Liu J, Wang S, Qian X, Lang M, Li J, Xie M, Li C, et al. 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. 2019;38:70–6.
28. Yin C, Ren Y, Adnan A, Tian J, Guo K, Xia M, He Z, Zhai D, Chen X, Wang L, et al. Developmental validation of Y-SNP pedigree tagging system: a panel via quick ARMS PCR. Forensic Sci Int Genet. 2020;46:102271.
29. Zhou Z, Zhou Y, Yao Y, Qian J, Liu B, Yang Q, Shao C, Li H, Sun K, Tang Q, et al. A 16-plex Y-SNP typing system based on allele-specific PCR for the genotyping of Chinese Y-chromosomal haplogroups. Leg Med (Tokyo). 2020;46:101720.
30. Claerhout S, Verstraete P, Warnez L, Vanpaemel S, Larmuseau M, Decorte R. CSYseq: the first Y-chromosome sequencing tool typing a large number of Y-SNPs and Y-STRs to unravel worldwide human population genetics. PLoS Genet. 2021;17(9):e1009758.
31. Tao R, Li M, Chai S, Xia R, Qu Y, Yuan C, Yang G, Dong X, Bian Y, Zhang S. Developmental validation of a 381 Y-chromosome SNP panel for haplogroup analysis in the Chinese populations. Forensic Sci Int Genet. 2023;62:102803.
32. Wang M, He G, Zou X, Liu J, Ye Z, Ming T, Du W, Wang Z, Hou Y. Genetic insights into the paternal admixture history of Chinese Mongolians via high-resolution customized Y-SNP SNaPshot panels. Forensic Sci Int Genet. 2021;54:102565.
33. Lang M, Liu H, Song F, Qiao X, Ye Y, Ren H, Li J, Huang J, Xie M, Chen S, et al. Forensic characteristics and genetic analysis of both 27 Y-STRs and 143 Y-SNPs in Eastern Han Chinese population. Forensic Sci Int Genet. 2019;42:e13–20.
34. Ralf A, van Oven M, Montiel Gonzalez D, de Knijff P, van der Beek K, Wootton S, Lagace R, Kayser M. Forensic Y-SNP analysis beyond SNaPshot: high-resolution Y-chromosomal haplogrouping from low quality and quantity DNA using Ion AmpliSeq and targeted massively parallel sequencing. Forensic Sci Int Genet. 2019;41:93–106.
35. Gao T, Yun L, Zhou D, Lang M, Wang Z, Qian X, Liu J, Hou Y. Next-generation sequencing of 74 Y-SNPs to construct a concise consensus phylogeny tree for Chinese population. Forensic Sci Int Genet Suppl Ser. 2017;6:e96–8.
36. Liu J, Jiang L, Zhao M, Du W, Wen Y, Li S, Zhang S, Fang F, Shen J, He G. Development and validation of a custom panel including 256 Y-SNPs for Chinese Y-chromosomal haplogroups dissection. Forensic Sci Int Genet. 2022;61:102786.
37. He GL, Wang MG, Zou X, Yeh HY, Liu CH, Liu C, Chen G, Wang CC. Extensive ethnolinguistic diversity at the crossroads of North China and South Siberia reflects multiple sources of genetic diversity. J Syst Evol. 2022;61(1):230–50.
38. He GL, Li YX, Zou X, Yeh HY, Tang RK, Wang PX, Bai JY, Yang XM, Wang Z, Guo JX, et al. Northern gene flow into southeastern East Asians inferred from genome-wide array genotyping. J Syst Evol. 2022;61(1):179–97.
39. Mao X, Zhang H, Qiao S, Liu Y, Chang F, Xie P, Zhang M, Wang T, Li M, Cao P, et al. The deep population history of northern East Asia from the Late Pleistocene to the Holocene. Cell. 2021;184(12):3256–66.
40. Wang T, Wang W, Xie G, Li Z, Fan X, Yang Q, Wu X, Cao P, Liu Y, Yang R, et al. Human population history at the crossroads of East and Southeast Asia since 11,000 years ago. Cell. 2021;184(14):3829–41.
41. Wang CC, Yeh HY, Popov AN, Zhang HQ, Matsumura H, Sirak K, Cheronet O, Kovalev A, Rohland N, Kim AM, et al. Genomic insights into the formation of human populations in East Asia. Nature. 2021;591(7850):413–9.
42. Yang MA, Fan X, Sun B, Chen C, Lang J, Ko YC, Tsang CH, Chiu H, Wang T, Bao Q, et al. Ancient DNA indicates human population shifts and admixture in northern and southern China. Science. 2020;369(6501):282–8.
43. Chen N, Ren L, Du L, Hou J, Mullin VE, Wu D, Zhao X, Li C, Huang J, Qi X, et al. Ancient genomes reveal tropical bovid species in the Tibetan Plateau contributed to the prevalence of hunting game until the late Neolithic. Proc Natl Acad Sci U S A. 2020;117(45):28150–9.
44. Chen J, He G, Ren Z, Wang Q, Liu Y, Zhang H, Yang M, Zhang H, Ji J, Zhao J, et al. Genomic insights into the admixture history of Mongolic- and Tungusic-speaking populations from southwestern east Asia. Front Genet. 2021;12(880):685285.
45. Wu Q, Cheng HZ, Sun N, Ma PC, Sun J, Yao HB, Xie YM, Li YL, Meng SL, Zhabagin M, et al. Phylogenetic analysis of the Y-chromosome haplogroup C2b–F1067, a dominant paternal lineage in Eastern Eurasia. J Hum Genet. 2020;65(10):823–9.
46. Zhang P, Luo H, Li Y, Wang Y, Wang J, Zheng Y, Niu Y, Shi Y, Zhou H, Song T, et al. NyuWa genome resource: a deep whole-genome

sequencing-based variation profile and reference panel for the Chinese population. Cell Rep. 2021;37(7):110017.

47. Cao Y, Li L, Xu M, Feng Z, Sun X, Lu J, Xu Y, Du P, Wang T, Hu R, et al. The ChinaMAP analytics of deep whole genome sequences in 10,588 individuals. Cell Res. 2020;30(9):717–31.
48. He GL, Wang MG, Li YX, Zou X, Yeh HY, Tang RK, Yang XM, Wang Z, Guo JX, Luo T, et al. Fine-scale north-to-south genetic admixture profile in Shaanxi Han Chinese revealed by genome-wide demographic history reconstruction. J Syst Evol. 2021;60(4):955–72.
49. McColl H, Racimo F, Vinner L, Demeter F, Gakuhari T, Moreno-Mayar JV, van Driem G, Gram Wilken U, Seguin-Orlando A, de la Fuente CC, et al. The prehistoric peopling of Southeast Asia. Science. 2018;361(6397):88–92.
50. Lipson M, Cheronet O, Mallick S, Rohland N, Oxenham M, Pietrusewsky M, Pryce TO, Willis A, Matsumura H, Buckley H, et al. Ancient genomes document multiple waves of migration in Southeast Asian prehistory. Science. 2018;361(6397):92–5.
51. Larena M, Sanchez-Quinto F, Sjodin P, McKenna J, Ebeo C, Reyes R, Casel O, Huang JY, Hagada KP, Guilay D, et al. Multiple migrations to the Philippines during the last 50,000 years. Proc Natl Acad Sci U S A. 2021;118(13):e2026132118.
52. Larena M, McKenna J, Sanchez-Quinto F, Bernhardsson C, Ebeo C, Reyes R, Casel O, Huang JY, Hagada KP, Guilay D, et al. Philippine Ayta possess the highest level of Denisovan ancestry in the world. Curr Biol. 2021;31(19):4219–30.
53. Kutanan W, Liu D, Kampuansai J, Srikummool M, Srithawong S, Shoocongdej R, Sangkhano S, Ruangchai S, Pittayaporn P, Arias L, et al. Reconstructing the human genetic history of mainland Southeast Asia: insights from genome-wide data from Thailand and Laos. Mol Biol Evol. 2021;38(8):3459–77.
54. Liu D, Duong NT, Ton ND, Van Phong N, Pakendorf B, Van Hai N, Stoneking M. Extensive ethnolinguistic diversity in Vietnam reflects multiple sources of genetic diversity. Mol Biol Evol. 2020;37(9):2503–19.
55. Macholdt E, Arias L, Duong NT, Ton ND, Van Phong N, Schröder R, Pakendorf B, Van Hai N, Stoneking M. The paternal and maternal genetic history of Vietnamese populations. Eur J Hum Genet. 2020;28(5):636–45.
56. Kutanan W, Shoocongdej R, Srikummool M, Hubner A, Suttipai T, Srithawong S, Kampuansai J, Stoneking M. Cultural variation impacts paternal and maternal genetic lineages of the Hmong-Mien and Sino-Tibetan groups from Thailand. Eur J Hum Genet. 2020;28(11):1563–79.
57. Wang J, Yang L, Duan S, Sun Q, Li Y, Wu J, Wu W, Wang Z, Liu Y, Tang R, et al. Genome-wide allele and haplotype-sharing patterns suggested one unique Hmong-Mein-related lineage and biological adaptation history in Southwest China. Hum Genomics. 2023;17(1):3.
58. Wang F, Song F, Song M, Luo H, Hou Y. Genetic structure and paternal admixture of the modern Chinese Zhuang population based on 37 Y-STRs and 233 Y-SNPs. Forensic Sci Int Genet. 2022;58:102681.
59. Song M, Wang Z, Zhang Y, Zhao C, Lang M, Xie M, Qian X, Wang M, Hou Y. 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. 2019;39:e14–20.
60. Song M, Wang Z, Lyu Q, Ying J, Wu Q, Jiang L, Wang F, Zhou Y, Song F, Luo H, et al. Paternal genetic structure of the Qiang ethnic group in China revealed by high-resolution Y-chromosome STRs and SNPs. Forensic Sci Int Genet. 2022;61:102774.
61. Huang YZ, Wei LH, Yan S, Wen SQ, Wang CC, Yang YJ, Wang LX, Lu Y, Zhang C, Xu SH, et al. Whole sequence analysis indicates a recent southern origin of Mongolian Y-chromosome C2c1a1a1-M407. Mol Genet Genomics. 2018;293(3):657–63.
62. Chen H, Lin R, Lu Y, Zhang R, Gao Y, He Y, Xu S. Tracing Bai-Yue ancestry in aboriginal Li people on Hainan Island. Mol Biol Evol. 2022;39:msac10.
63. He G, Wang Z, Guo J, Wang M, Zou X, Tang R, Liu J, Zhang H, Li Y, Hu R, et al. Inferring the population history of Tai-Kadai-speaking people and southernmost Han Chinese on Hainan Island by genome-wide array genotyping. Eur J Hum Genet. 2020;28(8):1111–23.
64. He G, Wang Z, Wang M, Luo T, Liu J, Zhou Y, Gao B, Hou Y. Forensic ancestry analysis in two Chinese minority populations using massively parallel sequencing of 165 ancestry-informative SNPs. Electrophoresis. 2018;39(21):2732–42.
65. Ma X, Yang W, Gao Y, Pan Y, Lu Y, Chen H, Lu D, Xu S. Genetic origins and sex-biased admixture of the Huis. Mol Biol Evol. 2021;38(9):3804–19.
66. He G, Fan ZQ, Zou X, Deng X, Yeh HY, Wang Z, Liu J, Xu Q, Chen L, Deng XH *et al*: Demographic model and biological adaptation inferred from the genome-wide single nucleotide polymorphism data reveal tripartite origins of southernmost Chinese Huis. Am J Biol Anthropol 2022, n/a(n/a).
67. Jeong C, Ozga AT, Witonsky DB, Malmstrom H, Edlund H, Hofman CA, Hagan RW, Jakobsson M, Lewis CM, Aldenderfer MS, et al. Long-term genetic stability and a high-altitude East Asian origin for the peoples of the high valleys of the Himalayan arc. Proc Natl Acad Sci U S A. 2016;113(27):7485–90.
68. He G, Wang M, Zou X, Chen P, Wang Z, Liu Y, Yao H, Wei LH, Tang R, Wang CC, et al. Peopling history of the Tibetan plateau and multiple waves of admixture of Tibetans inferred from both ancient and modern genome-wide data. Front Genet. 2021;12(1634):725243.
69. Liu Y, Xie J, Wang M, Liu C, Zhu J, Zou X, Li W, Wang L, Leng C, Xu Q, et al. Genomic insights into the population history and biological adaptation of southwestern Chinese Hmong-Mien people. Front Genet. 2021;12:815160.
70. Li L, Huang P, Sun X, Wang S, Xu M, Liu S, Feng Z, Zhang Q, Wang X, Zheng X, et al. The ChinaMAP reference panel for the accurate genotype imputation in Chinese populations. Cell Res. 2021;31(12):1308–10.
71. World Medical Association I: Declaration of Helsinki. Ethical principles for medical research involving human subjects. J Indian Med Assoc. 2009; 107(6):403–5.
72. Mallick S, Li H, Lipson M, Mathieson I, Gymrek M, Racimo F, Zhao M, Chennagiri N, Nordenfelt S, Tandon A, et al. The Simons genome diversity project: 300 genomes from 142 diverse populations. Nature. 2016;538(7624):201–6.
73. Pagani L, Lawson DJ, Jagoda E, Morseburg A, Eriksson A, Mitt M, Clemente F, Hudjashov G, DeGiorgio M, Saag L, et al. Genomic analyses inform on migration events during the peopling of Eurasia. Nature. 2016;538(7624):238–42.
74. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience. 2015;4:7.
75. Jagadeesan A, Ebenesersdottir SS, Guethmundsdottir VB, Thordardottir EL, Moore KHS, Helgason A. HaploGrouper: a generalized approach to haplogroup classification. Bioinformatics. 2021;37(4):570–2.
76. Chen H, Lu Y, Lu D, Xu S. Y-LineageTracker: a high-throughput analysis framework for Y-chromosomal next-generation sequencing data. BMC Bioinform. 2021;22(1):114.
77. Dellicour S, Gill MS, Faria NR, Rambaut A, Pybus OG, Suchard MA, Lemey P. Relax, keep walking—a practical guide to continuous phylogeographic inference with BEAST. Mol Biol Evol. 2021;38(8):3486–93.
78. Leigh JW, Bryant D, Nakagawa S. POPART: full-feature software for haplotype network construction. Methods Ecol Evol. 2015;6(9):1110–6.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.