BMC Biology

RESEARCH ARTICLE

Open Access

YHSeqY3000 panel captures all founding lineages in the Chinese paternal genomic diversity database

Mengge Wang[1,2,9,13*†], Shuhan Duan[1,2,3,13†], Qiuxia Sun[1,2,8], Kaijun Liu[4,5], Yan Liu[1,3], Zhiyong Wang[1,2,7], Xiangping Li[1,2,7], Lanhai Wei[6], Yunhui Liu[1,2,8], Shengjie Nie[7], Kun Zhou[5], 10K_CPGDP consortium[1*], Yongxin Ma[11], Huijun Yuan[1,2], Bing Liu[12], Lan Hu[12], Chao Liu[9,10*] and Guanglin He[1,2,9*]

Abstract

Background The advancements in second-/third-generation sequencing technologies, alongside computational innovations, have significantly enhanced our understanding of the genomic structure of Y-chromosomes and their unique phylogenetic characteristics. These researches, despite the challenges posed by the lack of population-scale genomic databases, have the potential to revolutionize our approach to high-resolution, population-specific Y-chromosome panels and databases for anthropological and forensic applications.

Objectives This study aimed to develop the highest-resolution Y-targeted sequencing panel, utilizing time-stamped, core phylogenetic informative mutations identified from high-coverage sequences in the YanHuang cohort. This panel is intended to provide a new tool for forensic complex pedigree search and paternal biogeographical ancestry inference, as well as explore the general patterns of the fine-scale paternal evolutionary history of ethnolinguistically diverse Chinese populations.

Results The sequencing performance of the East Asian-specific Y-chromosomal panel, including 2999-core SNP variants, was found to be robust and reliable. The YHSeqY3000 panel was designed to capture the genetic diversity of Chinese paternal lineages from 3500 years ago, identifying 408 terminal lineages in 2097 individuals across 41 genetically and geographically distinct populations. We identified a fine-scale paternal substructure that was correlating with ancient population migrations and expansions. New evidence was provided for extensive gene flow events between minority ethnic groups and Han Chinese people, based on the integrative Chinese Paternal Genomic Diversity Database.

Conclusions This work successfully integrated Y-chromosome-related basic genomic science with forensic and anthropological translational applications, emphasizing the necessity of comprehensively characterizing

†Mengge Wang and Shuhan Duan have contributed equally to this work and share first authorship.

*Correspondence:
Mengge Wang
Menggewang2021@163.com
Chao Liu
liuchaogzf@163.com
Guanglin He
guanglinhescu@163.com
10K_CPGDP consortium
Full list of author information is available at the end of the article

Y-chromosome genomic diversity from genomically under-representative populations. This is particularly important in the second phase of our population-specific medical or anthropological genomic cohorts, where dense sampling strategies are employed.

Keywords YanHuang cohort, Genomic diversity database, YHSeqY3000 panel, Forensic genomics, Human evolution

Background

Human genomic variations uncover the genetic foundations and evolutionary codes of diseases and traits within the complex interplay of genes and environmental factors [1, 2]. Ancient events, including demographic, cultural, and socioeconomic transitions, have shaped these variations. Patterns of genomic diversity and evolutionary trajectories reflect ancient population migrations and admixture events [3–5]. Despite advances in sequencing technologies and computational methodologies over the past two decades, the genomic structures of many previously unresolved complex regions and the overall landscape of genetic diversity in ethnolinguistically diverse human populations remain incompletely understood [6, 7]. Recent efforts by the Telomere-To-Telomere (T2T) consortium and Human Pangenome Reference Consortium have published a complete human reference genome sequence, elucidating genomic, epigenomic, and transcriptomic features of segmental duplications, tandem repeats, and acrocentric chromosomes [8–12]. The continuity of the T2T-CHM13 and graph-based reference genomes has enhanced capabilities in read mapping and the discovery of single nucleotide polymorphisms (SNPs) and structural variations (SVs), facilitating high-resolution examination of complex and clinically relevant genes [8]. However, this haplotype-based T2T genome has not fully resolved the complexities and variations of the human Y-chromosome. Rhie et al. assembled multiple T2T Y-chromosome genomes using PacBio HiFi and ONT Ultra-Long reads, revealing intricate phylogenetic features, complex patterns of tandem repeats, and segmental duplications [13]. Mutations in the Y-chromosome reflect the evolutionary processes of population admixture, expansion, and divergence, as these can be directly inherited from paternal ancestors [6]. Challenging regions of the Y-chromosome are strongly associated with human evolution and phenotypic diversity [14]. Techniques such as restriction fragment length polymorphisms (RFLPs), short tandem repeats (STRs), binary SNPs, and short insertions and deletions (InDels) genotyped via electrophoresis, denaturing high-performance liquid chromatography, or sequencing were prevalent in early research. Demic and cultural factors have significant influence on the patterns of genomic diversity of Y-chromosome genetic variations [6, 15–17]. Hammer et al. first investigated the Y Alu polymorphism among different continental populations, estimating divergence times and effective population sizes [18]. Their findings supported a recent common ancestor from Africa dating back approximately 188,000 years, rejecting the multiregional origin of *Homo sapiens* and the selection of favorable mutations in non-recombining Y-chromosome (NRY) regions [18]. This hypothesis provided the theoretical possibility for large-scale identifying continental populations' specific variations for precision paternal genetic history reconstruction and forensic applications. Cultural practices also significantly influence Y-chromosome genetic diversity patterns, including technology innovations, subsistence strategies, and marriage customs. Oota et al. examined the genetic polymorphisms of mtDNA HV1 sequences and Y-STRs in three patrilocal and three matrilocal hill tribes in Thailand, finding solid correlations between genetic diversity and residence patterns [19]. Patrilocal populations exhibited lower Y-chromosome diversity and higher mtDNA diversity compared to matrilocal societies. Gunnarsdóttir et al. found that matrilocal Semende from Sumatra had lower mtDNA diversity than patrilocal Besemah [20]. The genetic characteristics of NRY regions have since been widely utilized in forensics, medical genetics, human evolution studies, and genealogical reconstruction. Paternal genetic substructure among the ethnolinguistically Chinese population remained characterized via NRY regions.

The Y Chromosome Consortium (YCC) established a haplogroup nomenclature system to facilitate the universal and efficient use of binary variations in the male-specific Y-chromosome regions, creating a high-resolution human Y-chromosomal phylogeny tree [21]. Numerous studies have documented the genetic diversity of ethnolinguistically diverse populations and their migration events. Zerjal et al. examined the paternal genetic evolutionary history of northern East Asians and the Mongol Empire's expansion through Y-chromosome lineages [22]. Wen et al. analyzed phylogenetically informative SNPs from Y/mtDNA to assess how cultural diffusion and population movements influenced Han Chinese populations [23]. They concluded that demic diffusion significantly shaped the genetic makeup of the Han people [23]. Additional research using low-density mtDNA/Y-chromosome markers has shed light on the peopling process of the Tibetan Plateau and the expansion of Tibeto-Burman-speaking populations [24–28]. The accumulation of

Y-chromosome sequence data offers significant potential to enhance our understanding of human genetic history, benefiting a range of fields from evolutionary studies to forensic genetics [6, 7, 29–31]. Evidence supports the use of Y-chromosomal variants in forensic investigations, aiding in identifying paternal lineages, assisting in paternity tests, and inferring paternal biogeographic ancestry. Research has expanded our knowledge of the geographical distribution of dominant paternal lineages through Y-SNP analysis. For instance, van Oven et al. developed a multiplex SNP assay to dissect haplogroup O [32], and Yin et al. designed a pedigree tagging system with 24 Y-SNPs for East Asian haplogroup classification [33]. Our previous work created Y-SNP panels targeting haplogroups C, D, N, O, R, and others specific to Chinese ethnic groups using SNaPshot technology [34–36]. However, these tools are limited by the capillary electrophoresis system, which restricts the number of Y-SNPs that can be simultaneously genotyped, thus limiting the geographic, ethnic, and lineage resolution for biogeographic ancestry inference. To address these limitations, several next-generation sequencing (NGS)-based Y-SNP panels have been developed, allowing for the simultaneous typing of large numbers of Y-SNPs for more detailed paternal ancestry inference. Ralf et al. introduced a high-resolution multiplex tool for analyzing 530 Y-SNPs using targeted semiconductor sequencing [37]. They later improved this tool to simultaneously analyze 859 Y-SNPs, inferring 640 Y-chromosomal haplogroups [38]. In China, geneticists have developed Y-SNP NGS panels to classify paternal lineage among ethnolinguistically diverse groups. Gao et al. created a 74-plex Y-SNP NGS system, constructing a consensus phylogeny for Chinese populations [39]. This system was subsequently upgraded to 165-plex and 256-plex Y-SNP NGS panels based on previous Y-SNP SNaPshot assays [40, 41]. Tao et al. developed the SifaMPS 381 Y-SNP panel for the Chinese population haplogroup classification [42]. Recently, we also developed a higher-resolution 639-plex Y-SNP panel, genotyping and validating these SNPs in 1033 individuals from 33 diverse Chinese populations, which identified several population-specific founding lineages, revealing diversification and expansion patterns of different paternal lineages [43]. However, these studies still face limitations in lineage resolution and diversity coverage among East Asians.

Generally, previous Y-SNP screenings primarily relied on the ISOGG (International Society of Genetic Genealogy) and Yfull databases. However, the resolution and coverage of these public Y-chromosomal phylogenies were insufficient for the high-resolution paternal lineage classification required for ethnolinguistically diverse Chinese populations. Besides, Y-chromosome genomic resources are limited in human genome research, which limited our deep understanding of fine-scale paternal genetic structures. To address this, we established the Chinese Paternal Genomic Diversity Database (CPGDD) using the 10 K Chinese People Genomic Diversity Project (10K_CPGDP), YanHuang cohort (YHC) genomic resources, and other publically available genomic resources [6, 31, 44, 45]. Our goal was to identify high-resolution, population-specific, and lineage-focused markers for forensic and molecular applications. We developed the "YHSeqY3000" panel and genotyped 2999 panel-related Y-SNPs in 2097 male individuals from 41 diverse Chinese populations (Additional file 1: Table S1). By integrating our in-house genomic data with public datasets, we created a comprehensive and high-quality Y-chromosome database that was used to elucidate the impact of ancient population migrations, admixture, and agricultural innovations on the paternal genomic diversity of the Chinese population.

Results

Construction of the YHSeqY3000 panel

We initially identified 3010 Y-SNPs based on the carefully designed loci screening criteria and successfully designed primers for 3002 phylogenetically informative Y-SNPs. To validate the panel, we randomly selected 20 high-depth sequenced samples from the YHC for genotyping. Preliminary sequencing revealed that some Y-SNPs either lacked typing results or displayed heterozygous genotypes. These issues arose due to excessive filtering from multiple sequence alignments or complex sequences near the target Y-SNPs. To address this, we adjusted the filtering parameters, replaced low-quality loci with parallel Y-SNPs on the same branch, and redesigned primers. After resequencing with the improved Y-SNP panel, we found that variant information at three loci remained unavailable. Ultimately, we constructed a Y-SNP typing assay using the NGS platform, targeting 2850 defined NRY regions and covering 2999 paternal lineages (Additional file 2: Table S2). We named this assay the YHSeqY3000 panel.

Sequencing performance of the YHSeqY3000 panel

We used the YHSeqY3000 panel to sequence 20 randomly selected YHC samples, assessing genotyping consistency and repeatability. The variant genotypes matched those obtained through whole-genome sequencing. The average sequencing depth for these samples was 409.73X, with a range from 175.89X to 953.20X. Y-SNPs reaching Q30 constituted 94.68%, and the average mapping rate was 99.08%. The average capture rate was 65.99%, and more than 96.50% of target Y-SNPs had a sequencing depth greater than 30X (Fig. 1A). Additionally, all three sequencing runs produced complete and

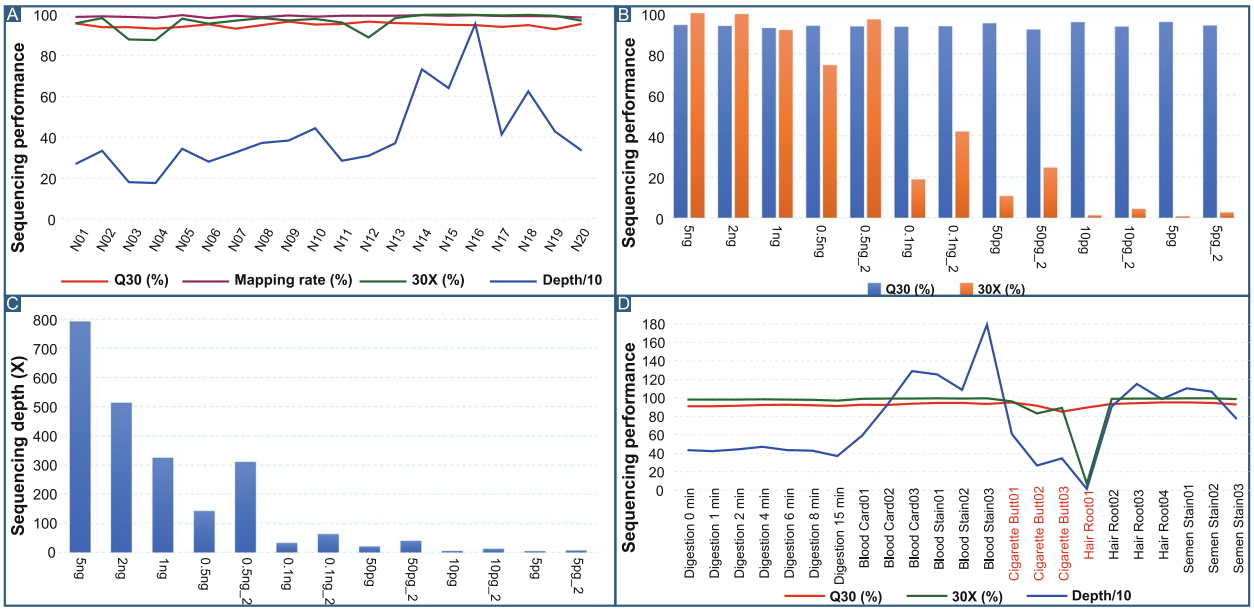

Fig. 1 Sequencing performance of the YHSeqY3000 panel. **A** Consistency testing results. The light blue line represents the average Q30 value per sample, the orange line represents the average mapping rate, the yellow line indicates the proportion of Y-SNPs exceeding 30X coverage, and the gray line represents one-tenth of the average sequencing depth. **B** Sensitivity testing results show the proportion of Y-SNPs with sequencing depth over 30X and base quality at Q30. **C** Average sequencing depth of control DNA 9948 at varying concentrations in sensitivity testing. **D** Stability testing results for enzyme-digested DNA standards and case-type samples. The solid lines' colors correspond to the parameters in **A**

consistent variant genotypes for control DNA 9948. To evaluate the sequencing sensitivity of the YHSeqY3000 panel, we tested varying amounts of control DNA 9948, ranging from 5 ng to 5 pg, in the initial amplification and adjusted the PCR cycles for low-input DNA during library construction. The sequencing results indicated that complete genotypes could be obtained with 5 to 1 ng of input control DNA 9948. At this range, the average sequencing depth varied from 792.63X to 324.99X. The proportions of Y-SNPs with sequencing depths above 30X were between 99.79 and 91.47%, and Y-SNPs achieving Q30 quality scores ranged from 93.90 to 92.46% (Fig. 1B–C). Increasing the number of PCR cycles in the second step of library construction from 9 to 12 significantly improved the average sequencing depth, the proportion of Y-SNPs exceeding 30X, and the proportion reaching Q30 (Fig. 1B–C). Overall, the results suggest that the optimal DNA input amount, under the recommended PCR cycles, is between 5 and 1 ng. For DNA input amounts less than 1 ng, increasing the PCR cycles can improve sequencing quality.

To validate the sequencing efficiency of the YHSeqY3000 panel on degraded DNA samples, we tested the stability of the DNase I enzyme by digesting the control DNA 9948 for varying degradation times. The results showed that the average sequencing depth exceeded 370.22X in all different degradation levels. Additionally, more than 97.05% of Y-SNPs had a coverage greater than 30X, and over 90.94% reached a quality score of Q30. Enzyme digestion for up to 15 min did not significantly impact variant calling (Fig. 1D). To evaluate the genotyping capability of the YHSeqY3000 panel on case-type samples, we sequenced various simulated case-type samples. After manually correcting the sequencing results using IGV, we found that complete genotypes were obtained for all samples except for one cigarette butt sample and one hair sample (Fig. 1D). To investigate the impact of PCR inhibitors on the performance of the YHSeqY3000 panel, we sequenced control DNA 9948 mixed with different concentrations of common inhibitors. Complete variant genotypes were achieved when the DNA standards were combined with less than 1.5 mmol/L EDTA, less than 10 mmol/L indigo, or less than 125 μM hemoglobin (Additional file 3: Table S3).

We validated the species specificity of the newly developed YHSeqY3000 panel using DNA from typical non-human samples. The results indicated that no reliable genotypes were obtained from any non-human DNA samples (Additional file 3: Table S3). Since mixture samples are common at forensic crime scenes, we performed sequencing on simulated male–female mixtures to evaluate the panel's genotyping ability on such samples. As the female component increased, the sequencing performance varied; however, all mixture samples yielded complete genotypes (Additional file 3: Table S3).

Genetic differentiation among linguistically related populations

China exhibits more linguistic diversity than other Eurasian regions, with language families such as Altaic (including Tungusic, Mongolic, and Turkic) prevalent in the northern of the Yellow River Basin. The Sino-Tibetan family, encompassing Sinitic and Tibeto-Burman languages, spans lowland China, the Tibetan Plateau, and nearby mid-altitude areas. In southern China, languages related to the Hmong-Mien, Austronesian, Austroasiatic, and Tai-Kadai families are linked to ancient rice-farming communities. The language-agriculture-people co-dispersal hypothesis has been proposed to explain these patterns of linguistic diversity and their relationship with ancient population movements and dynamic patterns of genomic diversity of spatiotemporally different people [46]. Recent linguistic phylogenies suggest that the origins, expansions, migrations, and divergences of ancient millet and rice farmers are associated with the Sino-Tibetan and Tai-Kadai language families [6]. Ancient DNA evidence from the Amur, West Liao, Yellow, and Yangtze River Basins, as well as the Tibetan Plateau, indicates that the origin and expansion of ancient Chinese source populations were pivotal in spreading their languages and cultures [46–52]. Additionally, ancient admixture patterns in northeastern Xinjiang and the Mongolian Plateau suggest that the Trans-Eurasian population and cultural exchanges influenced the genetic diversity of Neolithic peoples and their descendants [3–5, 53–56]. However, a systematic evaluation of paternal genetic patterns across geographically and linguistically diverse Chinese populations is still needed.

We genotyped dominant Y-chromosome mutations in 2097 individuals from 41 ethnolinguistically and geographically diverse Chinese populations and combined this data with previously published YHC data to investigate genetic continuity and admixture processes involving specific paternal lineages [31] (Fig. 2A–B). Our analysis identified 1673 unique haplotypes and 408 terminal paternal lineages (Additional file 4: Table S4). Haplotype diversity (HD) measures the likelihood that two randomly selected haplotypes from a population differ, serving as an indicator of genetic variation, with higher values reflecting greater diversity. Similarly, haplogroup diversity (HGD) quantifies the probability of two randomly chosen haplogroups differing, indicating the diversity of broader paternal lineages; both metrics help researchers analyze population structure and evolutionary history. We calculated HD to assess genetic similarity and differences among two randomly selected subjects and HGD to measure the similarity and differences in haplogroups among individuals based on haplogroup frequency (Additional file 4: Table S4). HD ranged from 0.9338 in the Kazakh population from Xinjiang to 1.0 in the Han from Yunnan, the Hui from Shandong, Hebei, Henan, Liaoning, the Manchu from Beijing, Hebei, Heilongjiang, Jilin, and Liaoning, the Li from Hainan, the Tibetan from Sichuan, as well as the Yao from Hunan. The HGD ranged from 0.7404 in the Kazakh population from Xinjiang to 1.0 in the Hui from Hebei, Henan, and Shandong, and the Manchu from Hebei, Heilongjiang, and Jilin. Estimates from discrimination probability (DP) and match probability (MP) also confirmed the high efficiency of these panels. Our estimates suggested a high genetic diversity of Chinese populations.

We reconstructed the phylogenetic tree using a merged Y-chromosome database and found that the O-P186 lineages and their sublineages were the predominant paternal lineages of ancient East Asians (Fig. 2C). The sublineages of O2a1b1a2, O2a1b1a1a1, O2a2b1a2a1a1b, O2a2b1a1a1a4, O2a2b1a1a1a1, O2a2b1a1a1a3, O2a2b1a1a1c, O2a2, and O2a2a1 from O2, as well as O1a1a2, O1a1a1a1a1, O1b1a2a, O1b1a1a1a1a1, and O1b1a1a1a1b1 from O1, were key founding lineages in the formation of the Chinese O-related gene pool. Additionally, founding lineages related to Eurasian steppe herders or Mongolian Plateau hunter-gatherers, such as R1b1, Q1a1a1, C2b1a2, C2a1, and Tibetan-related D1a1 lineages, also constituted a significant portion of the Chinese paternal gene pool. We further analyzed the network-based phylogenetic relationships to explore the evolutionary connections of paternal lineages and the population composition of the dominant lineages (Fig. 2D–E). We confirmed the occurrence of population expansion events in the phylogeny, observing several star-like expansions (such as O1) within these founding lineages (Fig. 2D). East Asian populations maintained long-term genetic stability or continuity in the Amur River Basin, Tibetan Plateau, and Yellow River Basin, which are centers of agricultural origin [47, 48]. We hypothesized that ancient population migration and admixture played limited roles in reshaping the genetic diversity patterns of ancient and modern Chinese populations. We anticipated that the founding lineages we identified would dominate ethnicity- or geography-specific populations. Our analysis revealed that several dominant paternal lineages, such as those mentioned above O-related farmer-related lineages from network results, contributed to the formation of multiple ethnolinguistically diverse modern populations, indicating that ongoing population migrations and interactions enriched the patterns of genetic admixture (Fig. 2E).

To investigate the genetic relationships between the populations in our study and other Chinese reference populations from the YHC, we conducted a series of principal component analyses (PCAs) using the top three components extracted from various population sets of different sample sizes (Fig. 3A; Additional file 5: Fig. S1). Our analysis revealed that Turkic-speaking Uyghur and

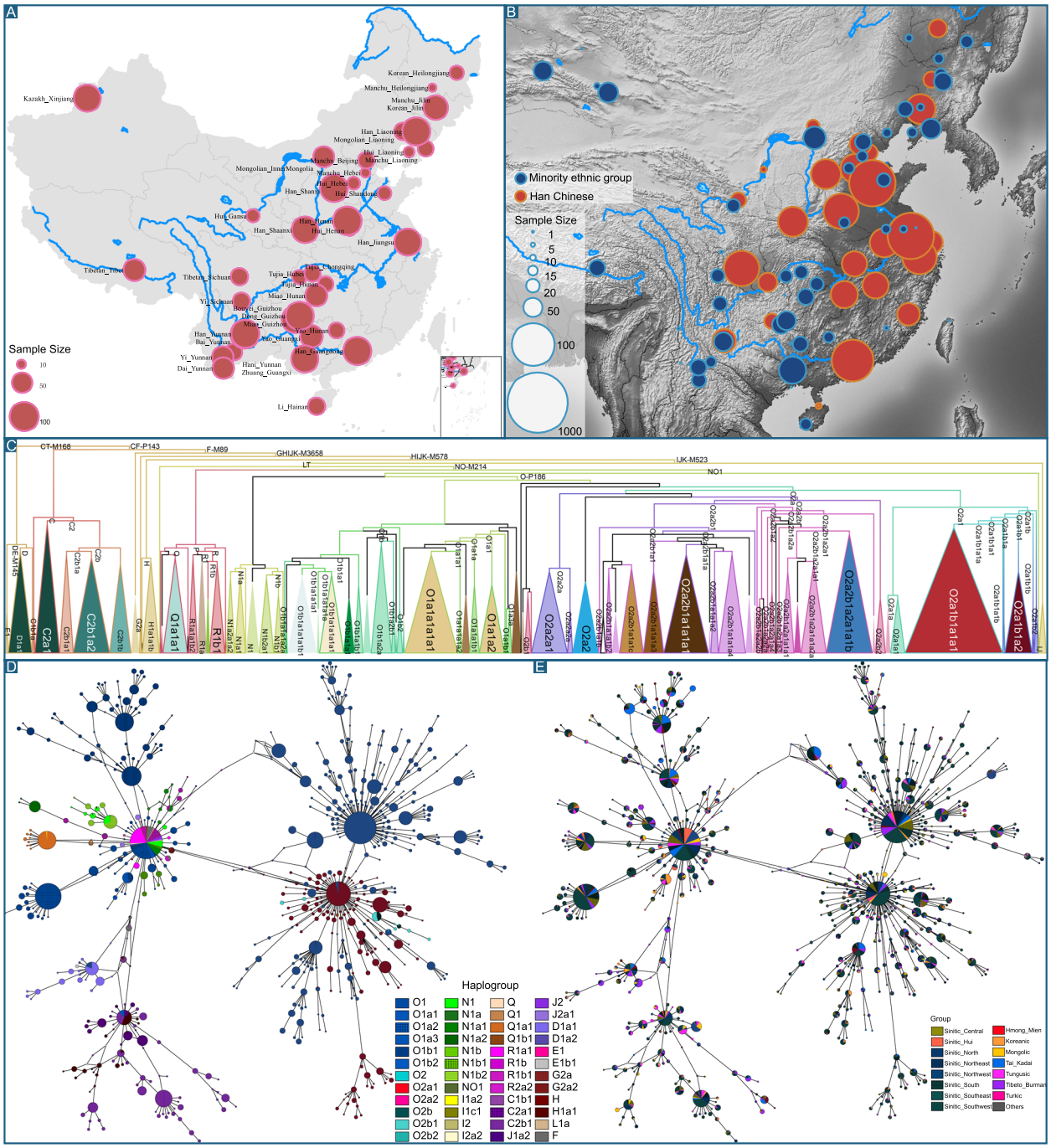

Fig. 2 Geographical distribution, population size, phylogeny, and network relationships of collected samples and YanHuang cohort phase 1 reference samples. **A** Geographical positions of 41 newly collected Chinese populations, representing 18 minority ethnic groups. Circle size corresponds to sample size. **B** Locations of 8036 individuals in the merged dataset, including 2375 minority ethnic individuals (dark blue) and 5661 Han Chinese from 29 populations (dark red). Circle size reflects population size. **C** A phylogenetic tree of 8036 targeted Y chromosome sequences was reconstructed via BEAST. The main lineages are labeled on the Y-chromosome phylogenetic tree. **D**–**E** Network analysis estimated admixture and expansion events of ancient paternal lineages. Colors represent haplogroup lineages (**D**) and geographical origin (**E**)

Kazakh populations were distinct from other groups at different levels of terminal paternal lineages, reflecting the relatively high proportion of R haplogroups among northwestern Chinese ethnic minorities (Fig. 3A). Similarly, other northwestern populations, such as the Tungusic-speaking Xibe, displayed genetic differentiation

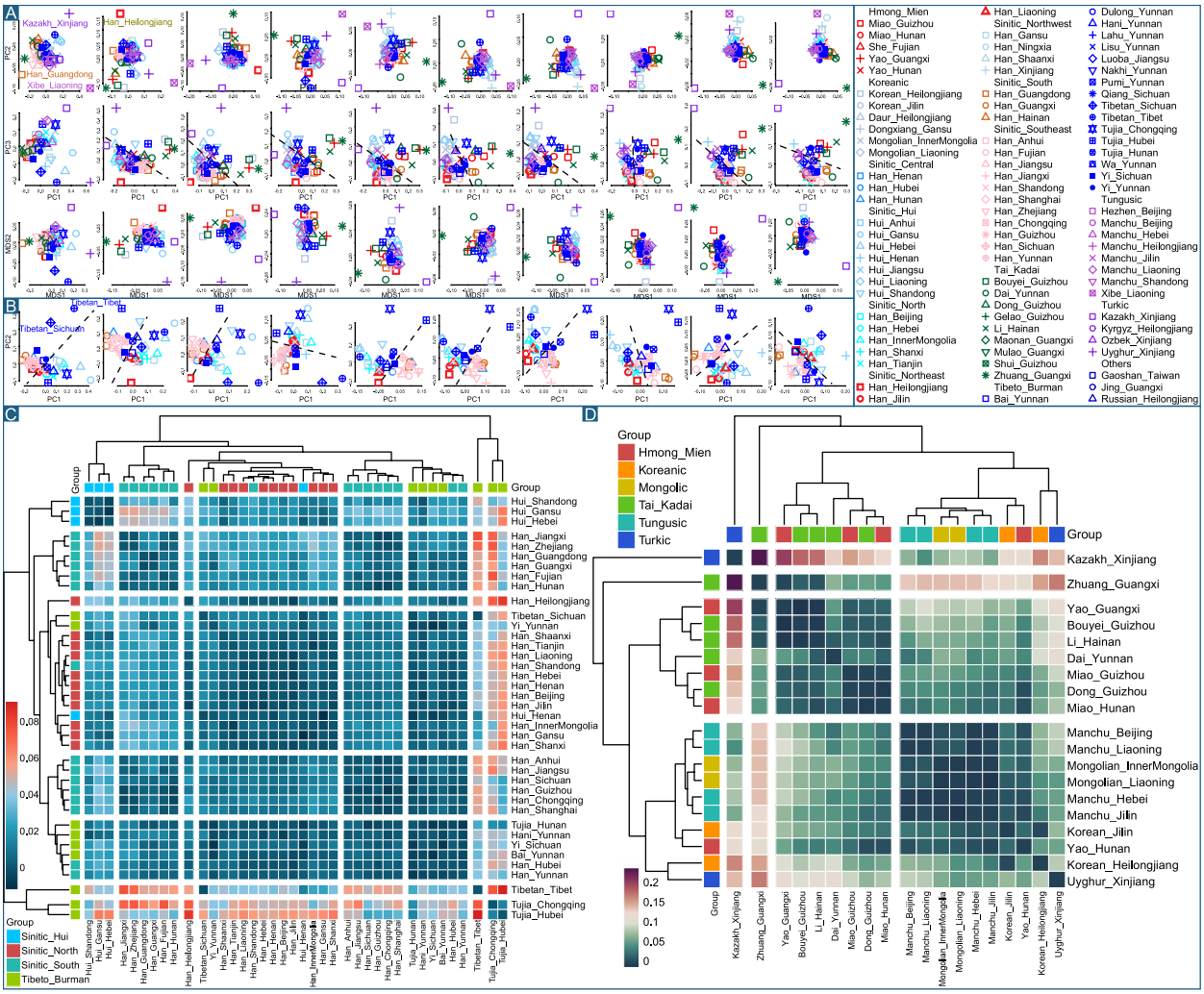

Fig. 3 Genetic structure of 88 ethnolinguistically diverse Chinese populations. **A** Principal component analysis (PCA) and multidimensional scaling (MDS) plots illustrate genetic similarities and differences among 88 Chinese populations. Different colors represent linguistically diverse Chinese populations, while shapes in similar colors denote distinct populations within a linguistic family. PCA was constructed using the top three components, and MDS was based on the Fst matrix. **B** PCA results for Sino-Tibetan-speaking populations. **C** The heatmap of pairwise Fst among 6317 individuals from 39 Sino-Tibetan-speaking populations shows genetic differentiation between Tibeto-Burman speakers and Sinitic people and between northern and southern Han Chinese. Red indicates larger Fst values and greater genetic differences, while blue denotes smaller genetic distances and stronger genetic affinity. **D** Genetic differences and affinities among 1509 individuals from 19 linguistically diverse Hmong-Mien-, Tai-Kadai-, and Altaic-speaking populations

from other groups. In northeastern China, populations exhibited a high proportion of C/Q-related lineages, supporting this region's role as the origin center of the Proto-Tungusic language and ancient northeastern Asians linked to DevilsCave, Boisman, and Neolithic Amur people [6]. We found that Han Chinese and Koreans from Heilongjiang province were genetically distinct from other groups, likely due to their C-related lineages, which were most frequent in northeastern Han Chinese populations (Fig. 2E). The Hui people's genetic profile indicated recent connections with Central and South Asians, reflecting historical trade along the Silk Road. PC3 analysis separated the Hui from other reference populations, particularly those from Gansu, Hebei, Shandong, and Henan provinces (Fig. 3A). Highland Tibetan people have a high proportion of D-related ancestral or descendant Y-chromosome lineages, which display a different frequency spectrum than those found in other East Asian populations (Fig. 4). Previous genetic studies have highlighted deep connections between Tibetans, Japanese, and Andamanese through shared D-related lineages [28]. We also found that Tibeto-Burman-speaking Tibetan

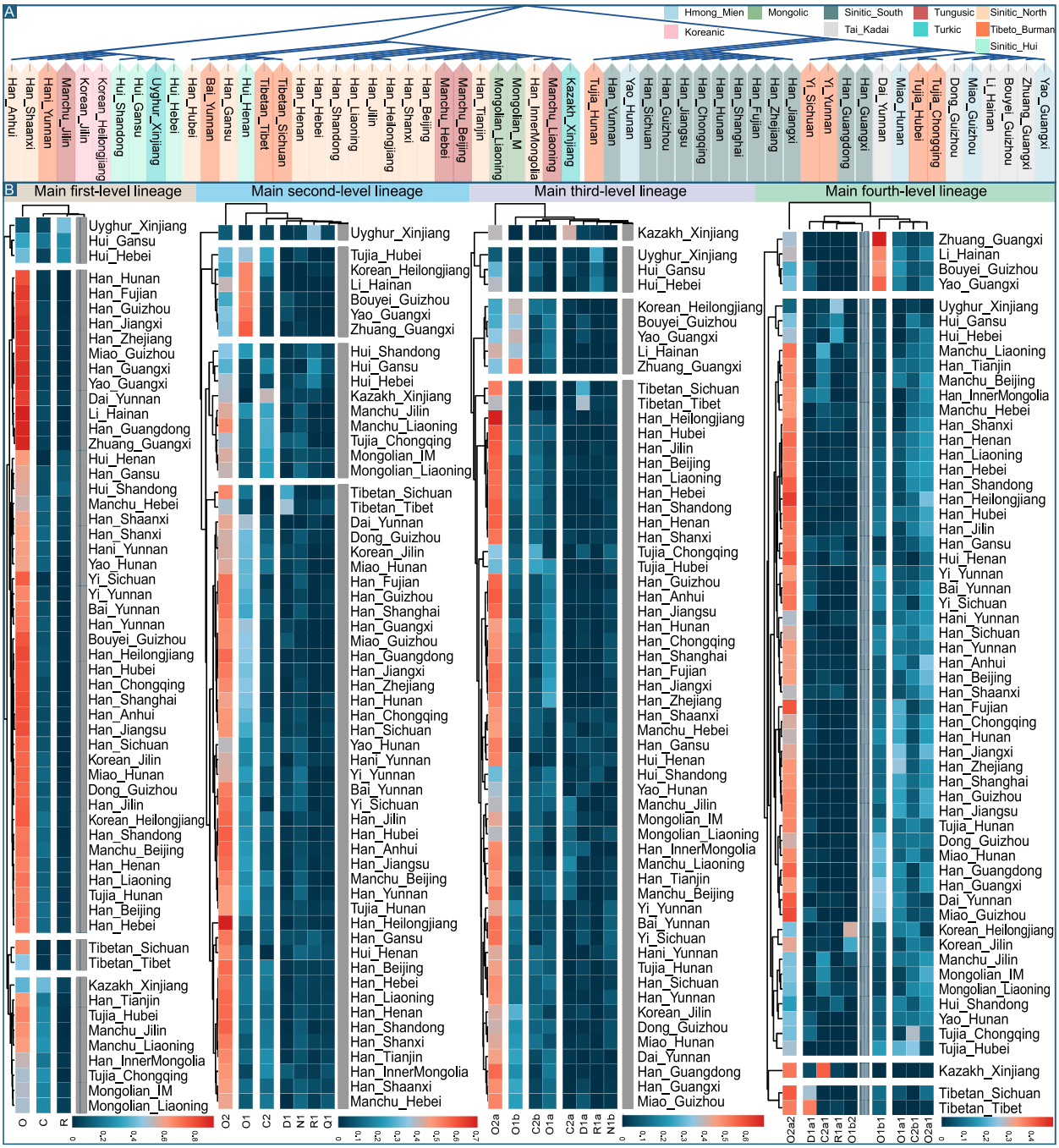

Fig. 4 Phylogenetic structure and haplogroup frequency spectrum of Y-chromosome lineages among 58 populations. **A** The neighbor-joining tree for 58 Chinese populations, reconstructed using Fst genetic distances, shows different background colors representing distinct language families. **B** Haplogroup frequency of dominant Chinese lineages across 58 populations, with red indicating significant frequencies and blue indicating minor frequencies

and Tujia populations also exhibited distinct clustering patterns. The Tai-Kadai-speaking Dai and Zhuang populations in Yunnan, southwest China, showed isolated clustering patterns compared to other groups. Multidimensional scaling (MDS) plots further confirmed these population affinities, showing consistent genetic relationships with those observed in the PCA, highlighting the genetic differentiation among geographically and linguistically diverse Chinese populations (Fig. 3A).

To explore the fine-scale population substructure and genetic relationships among linguistically similar but geographically distinct populations, we focused on Sino-Tibetan speakers, who have a large population size and are widely distributed in China and worldwide. The PCA clustering patterns revealed that Tibeto-Burman Tujia, Tibetan, and Yi speakers are distinct from others (Fig. 3B). We also observed a clear distinction between northern and southern Han Chinese populations. A heatmap of pairwise Fst indexes showed that the Hui people, along with Tibeto-Burman-speaking Tibetan and Tujia, had large genetic distances from other Sino-Tibetan reference populations (Fig. 3C). Additionally, there was clear genetic differentiation between northern and southern Han Chinese, with southern Han Chinese closely related to southern Tibeto-Burman speakers. We confirmed the genetic differentiation between northern Altaic-speaking populations and southern Chinese Hmong-Mien and Tai-Kadai people (Fig. 3D). A neighbor-joining tree revealed genetic similarities and differences, with two main branches: the northern branch included northern Han Chinese and northern Tibeto-Burman-speaking populations, while the southern branch encompassed southern Han Chinese, Tai-Kadai, and Hmong-Mien speakers (Fig. 4A). Our population genetic results suggested that fine-scale Chinese paternal genetic structures are associated with geographical and cultural divisions.

The effect of differentiated founding lineages on paternal genomic diversity and population differentiation

Bottleneck events observed in Y-chromosome diversity from high-coverage sequences suggest that cultural changes influenced the human out-of-Africa event and Neolithic male population size contraction. Further exploration is needed to understand how ancient human demographic histories affected Chinese paternal genetic diversity. Previous research has shown that lineages O, C, D, Q, and R are major components of China's paternal gene pool [6]. We explored the patterns of genetic diversity and genetic composition in our studied populations and identified additional second-level lineages (O2, O1, C2, D1, N1, R1, and Q1), third-level lineages (O2a, O1b, C2b, O1a, C2a, D1a, R1a, and N1b), and fourth-level lineages (O2a2, D1a1, C2a1, R1a1, O1b2, O1b1, O1a1, C2b1, and O2a1) as significant contributors to genetic differences between linguistically and geographically diverse populations (Fig. 4B). We estimated the correlation coefficient between geographical coordinates (latitude and longitude) and genetic differentiation indexes (Fst matrices among geographically different Han Chinese and genetic differences between Chinese minority ethnic groups). A negative correlation was observed between latitude and northern Han-related Fst values, and a positive correlation was observed with genetic differences in southern Han Chinese populations. These patterns are consistent with the close genetic affinities between southern Han Chinese and southern reference populations and the large genetic distances between northern Han Chinese and southern Chinese reference populations (Fig. 5A). No significant correlations were found between Fst genetic distances and latitude for three central Han Chinese populations (Anhui, Jiangsu, and Hubei), nor between pairwise Fst values of all Han Chinese groups, except Sichuan Han and longitude. This suggests substantial paternal genetic differences between northern and southern Han populations but no apparent differences between western and eastern Han Chinese. We also explored the correlations between longitude/latitude and genetic differentiation among minority ethnic groups. Latitude was significantly correlated with genetic differences between southern Chinese minority groups (Yi and Dai from Yunnan, Yao and Zhuang from Guangxi, Dong, Bouyei, and Miao from Guizhou, Miao from Hunan, Tujia from Hubei, and Li from Hainan) and other reference populations (Fig. 5B). Pairwise Fst values between northern Altaic-speaking populations (Kazakh and Uyghur from Xinjiang, Mongolian from Liaoning and Inner Mongolia, and Manchu from Jilin), as well as Hui from Gansu and other reference populations, showed a negative correlation with latitude. Interestingly, a negative correlation between genetic differentiation and longitude was identified for populations such as Mongolian from Inner Mongolia and Liaoning, Manchu from Jilin, Korean from Jilin and Heilongjiang, Hui from Shandong, Hani from Yunnan, Yao and Tujia from Hunan, and Dong from Guizhou. These correlation patterns suggest north-to-south and west-to-east genetic differentiation among minority ethnic groups. The genetic distance inter-correlations show northern and southern clusters among Han Chinese and genetic correlation gradients among minority ethnic groups.

We investigated which founding lineages contributed to the genetic differences among geographically and linguistically diverse populations, focusing on the correlation between fifth-level founding lineages and genetic differentiations (Fig. 5). The haplogroup frequencies of O2a2a and N1b2a showed a positive correlation with genetic differentiation estimates between northern Han and southern Han populations, and those of O1b1a and O1a1a had a negative correlation in the opposite direction. This suggests that these four founding lineages played distinct roles in shaping the paternal genetic makeup of northern and southern Han Chinese. We also identified differentiated correlations involving the lineages O1b2a, R1b1a, C2a1a, R1a1a, N1a1a, C2b1a, Q1a1a, and O2a2b with northern and southern genetic differentiation (Fig. 5A).

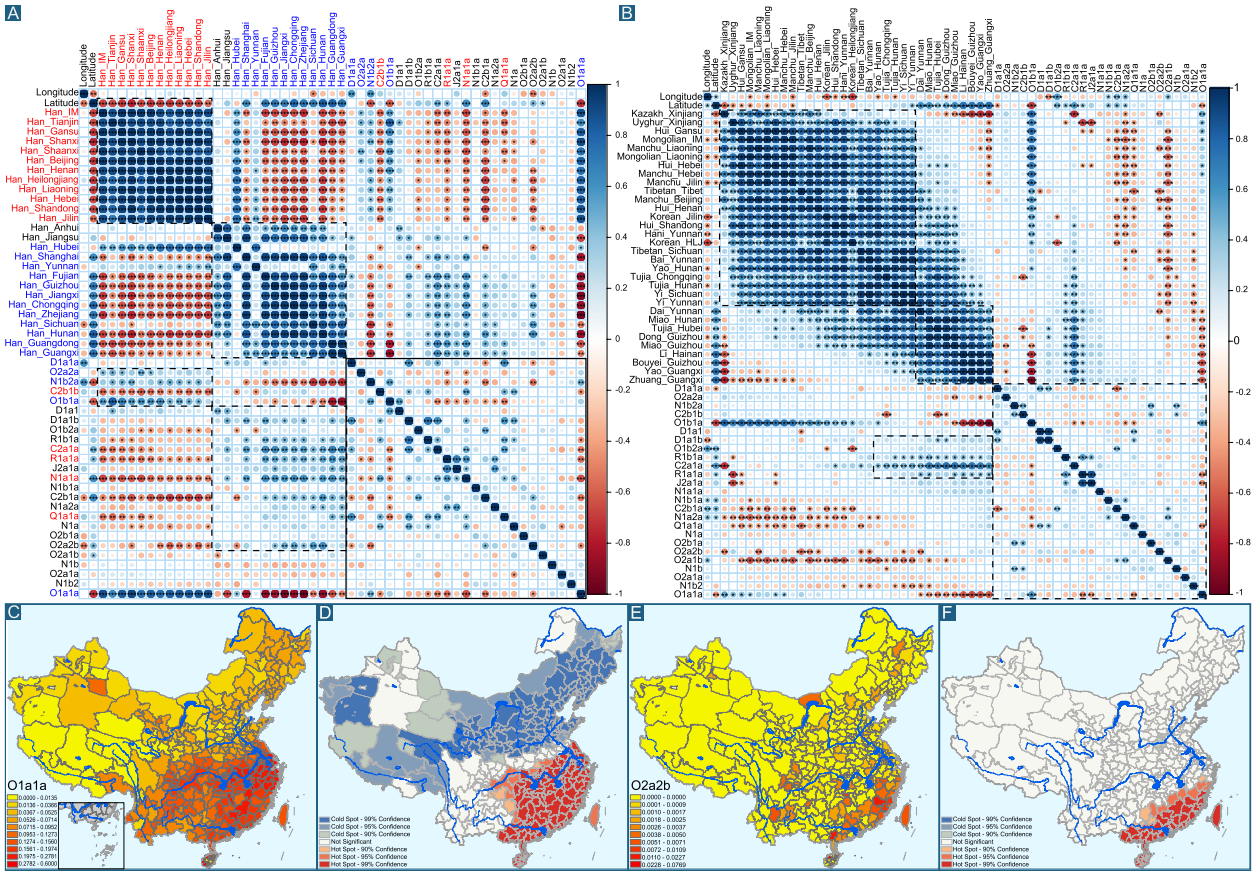

Fig. 5 Patterns of paternal genetic structure in Chinese populations. **A** Correlations between Fst genetic distances among 29 Han Chinese populations and lineage frequencies across 25 haplogroups. It also illustrates the correlation between Fst genetic distances and lineage frequencies among geographically distinct Han populations. The heatmap top displays correlations among geographical coordinates, genetic difference indexes, and lineage frequencies. **B** Correlations between geographical coordinates, genetic differentiation indexes, and lineage frequencies among 32 minority ethnic groups and 25 paternal lineages. Three stars indicate *p* values greater than 0.05, two stars denote *p* values from 0.01 to 0.05, and one star represents *p* values from 0.001 to 0.01. **C**–**F** Phylogeographical distribution and hotspot analysis of putatively southern East Asian-related founding lineages

Distinct correlation patterns between lineage frequency and genetic differentiation were notably observed in O1b1a, R1b1a, and C2a1a among northern and southern linguistically diverse minority ethnic groups (Fig. 5B).

We finally investigated how different founding lineages affect genetic differentiation in Chinese Y-chromosome diversity. We used a high-resolution phylogeographical analysis based on our in-house database, which integrates data from the 10K_CPGDP, YHC, and publicly available sources [6, 31, 55]. Our analysis identified two lineages with high frequencies in South China, possibly linked to the origins and spread of rice farmers and their proto-languages. The O1a1a lineage, with frequencies ranging from 0 to 0.6, is broadly distributed across the Yangtze and Zhujiang Basins. Spatial auto-correlation suggests that this founder lineage originated in the southeastern coastal regions (Fig. 5C–D). Another lineage, present at low frequency across China but peaking at ~ 0.0769 in Fujian and Taiwan, is estimated to have originated in the coastal regions of Hainan, Fujian, and Guangdong provinces (Fig. 5E–F). In northern East Asians, we observed distinct frequency patterns for dominant lineages. O2a1a, O2a2a, and D1a1b show high frequencies among northern East Asians and Tibetan Plateau populations (Fig. 6A–F), while Q2a2b, C2b1b, and R1a2b are most frequent at the crossroads between north China and Siberia or Central Asia (Fig. 6G–L). O2a1a frequencies range from 0 to 0.0377 among city-level Chinese populations, with the highest frequencies in northeastern, northern, and eastern China, including Qinghai and Gansu. Its estimated origin center is in the coastal regions of the lower Yangtze and Yellow Rivers. O2a2a exhibits similar frequency distributions and origin centers. D1a1b shows the highest frequency, over 0.1818, among highland East Asians, with an origin center near the Tibetan Plateau. Q2a2b peaks in northern China,

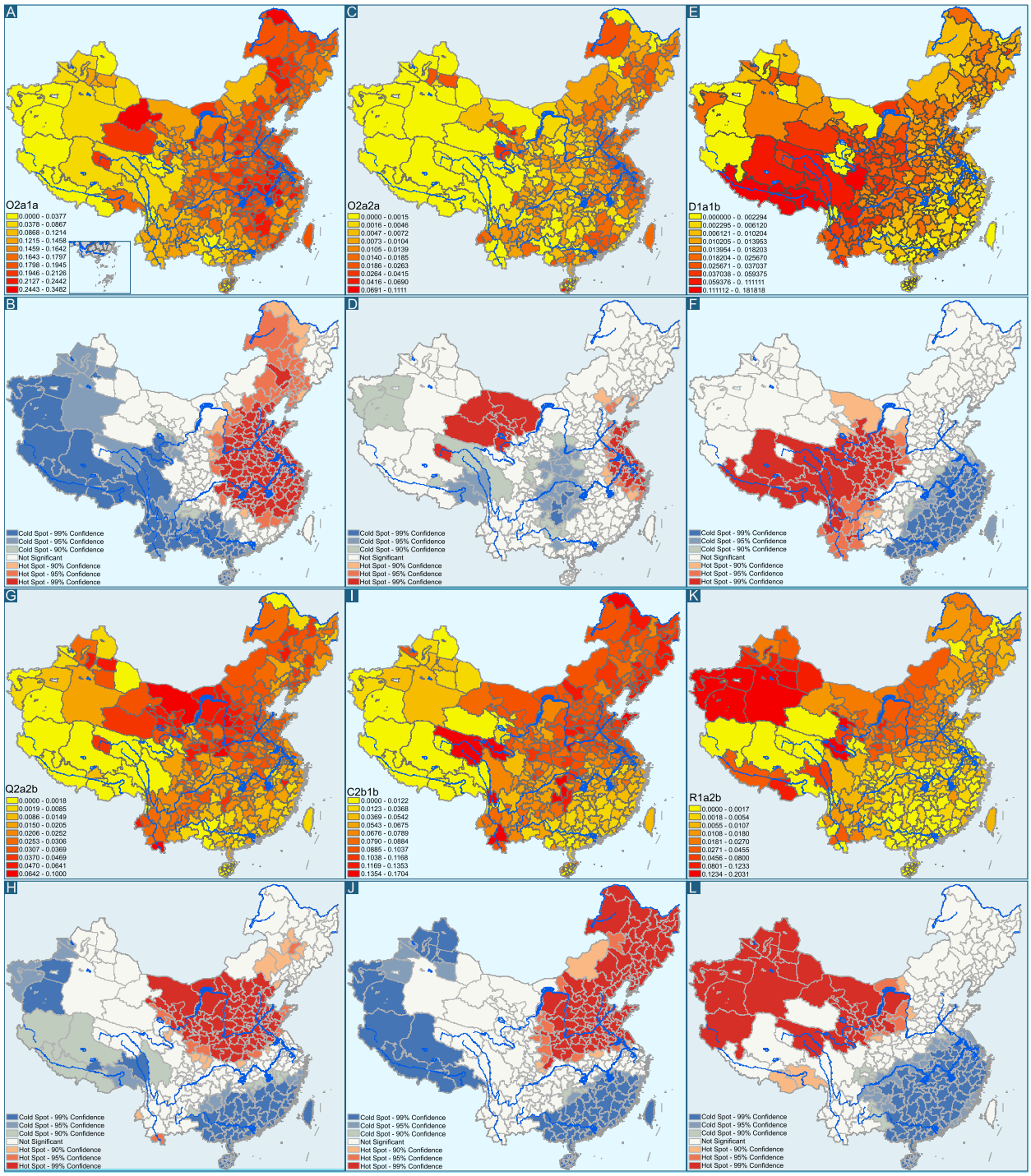

Fig. 6 Impact of Y-chromosome genomic diversity on genetic differentiation in northern and southern East Asians. **A**–**D** Geographical distribution and estimated hot spot patterns of specific founding lineages. **E**–**F** Haplogroup frequency of early Tibetan-related lineages and their potential origin center. **G**–**L** Geographical origin and hot spot analysis of C, Q, and R sublineages

particularly in Inner Mongolia, Shaanxi, and Shanxi, with its phylogenetic origin in the middle Yellow River Basin. C2b1b is widely distributed in northeastern China, with the highest frequencies in Heilongjiang, Jining, Liaoning, and Inner Mongolia; its geographical origin is estimated to be in northeastern China. High frequencies of C2b1b

in some southern Chinese populations may have resulted from migration and admixture during the Yuan and Ming dynasties. R1a2b is most frequent in northwestern and northern China, with its origin centered around Xinjiang, western Inner Mongolia, Gansu, and Qinghai provinces. Our findings suggest that the genetic differentiation between northern and southern Chinese paternal lineages was influenced by extensive ancient population expansions associated with millet and rice farming and trans-Eurasian population interactions.

Discussion

The genomic diversity of Y-chromosomes has been underrepresented in studies of human genome variation due to the complexity of sequencing and assembling regions enriched with segmental duplications, long palindromes, and tandem repeats. Furthermore, there is a lack of large-scale population cohorts focused on elucidating these evolutionary features [5, 6]. This study developed one high-resolution Y-chromosome SNP panel and built comprehensive paternal genomic resources to characterize Chinese Y-chromosome genetic diversity and evolutionary history. Our following population genetics also identified fine-scale geography and ethnicity-related paternal genetic substructure and illuminated that large-scale population genomics from under-representative populations was essential for human genetics, forensic science, and molecular anthropology. Previous analyses using low-resolution genotyping techniques, such as SNaPshot and pyrosequencing, have laid the foundation for understanding the origin of modern humans in Africa and their migration out of Africa approximately 50,000 years ago [57]. Y-chromosome genetic variations are crucial for inferring subsequent human migration and admixture events among non-African populations [22]. In Europe, genetic studies of Y-chromosome variations have highlighted the complex genetic influences of Near East farmers and Western Eurasian steppe herders [58]. Recent analyses of ancient autosomal DNA have supported an admixture model involving three ancestral sources contributing to the genomic formation of modern Europeans [59]. Y-chromosome data have also revealed early genetic connections between Asians and Oceanians and a Neolithic link between East Asian and Oceanian Austronesian-speaking populations [60]. Paternal genetic evidence from Chinese populations has significantly enhanced our understanding of early human settlement in Asia. The diverse Y-chromosome variations observed on the Tibetan Plateau suggest that both Paleolithic colonization and Neolithic expansion contributed to the genetic makeup of ancient and modern Tibetans [27]. Wen et al. demonstrated that demic diffusion patterns were crucial in shaping the paternal gene pool of northern and southern East Asians, with male-dominant migrations and sex-biased admixture adding complexity to the genetic structure of southern Chinese populations [23]. Zerjal et al. studied Y-chromosome variations in Altaic-speaking populations, such as the Xibe, highlighting the Mongol Empire's significant impact on the genetic structure of geographically diverse Eurasian populations [22]. They also showed that the Xibe people in the Xinjiang Uyghur Autonomous Region have close genetic ties with Tungusic speakers in the Amur River Basin.

This work presents a fine-scale analysis of paternal genetic structure, combined with high-resolution Y-chromosome phylogenetic topology and databases, which form the foundation for forensic science and confirmed the complex patterns of genetic diversity observed in population genetic, and anthropological applications based on autosomal evidences [3, 4, 7, 45]. Recent forensic studies have provided limited insights into the genetic structure of specific paternal lineages using advanced genotyping or sequencing panels. Song et al. investigated the paternal genetic structure of the Li people from Hainan Island, identifying O1-related founding lineages as the dominant Y-chromosome components in Tai-Kadai-speaking populations [61]. Subsequent studies on Han Chinese, Zhuang, Qiang, Hui, Tibetan, and Mongolian populations have furthered our understanding of forensic applications and the paternal genetic architecture of geographically diverse Chinese populations [35, 36, 41, 62, 63]. Recent whole-genome Y-chromosome sequencing has provided a detailed phylogenetic tree to shed light on human evolution. This research has revealed complex population expansion, migration, and admixture events among Oceanian people and their Neolithic connections with East Asians, based on Y-chromosome sequencing data and phylogeny [60]. Studies on whole Y-chromosome genomes from Siberian and American individuals have inferred a Beringian strait standstill and the founding lineages of Native Americans, marked by rapid population expansion [64]. Whole Y-chromosome sequencing data from Chinese populations, although with small sample sizes, emphasized the C, Q, D, and O lineages. These studies highlight the contributions of these lineages to the gene pool of northern Han Chinese, the peopling of the Tibetan Plateau, and the shared ancestry of southern Han Chinese and Tai-Kadai-speaking populations [65, 66].

We presented the large-scale Y-chromosome genomic resources for understanding the paternal genomic diversity of under-representative groups and patterns of Chinese fine-scale population structure. Although some advances in the Y-chromosome population genetic research, we still have some limitations compared with autosome-based work. Recent advances in Chinese

genome sequencing projects, such as the STROMICS genome study, non-small cell lung cancer cohort, Westlake BioBank for Chinese, NyuWa Genome resource, CHN100K, 10K_CPGDP, and China Metabolic Analytics Project, have been facilitated by reduced sequencing costs and rapid innovations in sequencing and computational statistics [55, 67–72]. However, these projects have limited information on the Y-chromosome, focusing primarily on human evolution and demographic modeling. We previously developed a high-resolution panel with 639 Y-chromosome SNPs to explore the paternal genetic structure of Chinese Mongolian, Gelao, Li, and Han populations [43, 73]. Despite its utility, this panel is limited in its ability to discriminate young and terminal Y-chromosome lineages, particularly among non-Han Chinese minority ethnic groups. To address these limitations, we initiated large-scale human genomic diversity projects, such as 10K_CPGDP and YHC, to reconstruct population evolutionary history using autosomal allele sharing, haplotype fragments, and uniparental lineages from mtDNA and the Y-chromosome [6, 31]. We developed a high-resolution Y-chromosome panel using Chinese-specific phylogenetic informative SNPs and a Y-chromosome phylogenetic topology reconstructed from our in-house integrated human genomic database. We validated the panel's sequencing performance by assessing sensitivity, specificity, concordance, repeatability, stability, and effectiveness with various case types, mixtures, and degraded samples. Our findings demonstrated the YHSeqY3000 panel's superior sequencing performance. This study provides new genomic data from underrepresented populations, addressing the gap in genetic diversity among ethnolinguistically diverse Chinese populations. We reconstructed the network-based phylogenetic topology and identified population expansion events in certain founding lineages, aligning with the Neolithic expansion of millet farmers from the Yellow River Basin and rice farmers from the Yangtze River Basin. Genetic analysis of populations from the YHC phase 1 revealed genetic differentiation among linguistically diverse minority ethnic groups. Altaic-speaking populations showed greater genetic distances from other Chinese reference populations than other population pairs, and the Fst-matrix correlated with longitudinal changes, suggesting genetic influences from Western Eurasian populations (Figs. 2, 3, and 4). Archeological evidence has revealed significant Bronze Age elements related to Western Eurasia—such as bronze, horse, sheep, barley, and wheat—at northwestern Chinese sites. Linguistic studies have also identified Indo-European-related language elements in the ancient Tianshan regions. Yao et al. examined the genetic structure of northwestern Han Chinese populations and found notable influences of Western Eurasian ancestry [53, 54]. Similarly, Wang et al. observed genetic contributions from Central Asians to ethnolinguistically diverse populations in the Hexi Corridor [55]. Genomic studies on Hui and Uyghur populations in Northwest China reported approximately 50% Western Eurasian ancestry in Uyghurs and found that differentiated Eurasian genetic variations significantly influenced phenotypic diversity, biological adaptation, allele-specific gene expression, and various quantitative trait locus functions [74, 75]. We conclude that population admixture has served as a powerful evolutionary force, contributing to the genomic and phenotypic diversity of northwestern Chinese populations, including Mongolic/Turkic speakers, on both autosomal and Y-chromosome levels.

A significant advancement of this study is the provision of a comprehensive lineage frequency spectrum and the identification of the potential origin centers of specific founding lineages in the CPGDD database (Figs. 5 and 6). Our findings suggest that the R1a2b-related western Eurasian lineages played a crucial role in the genetic differentiation between northwestern Chinese populations and other reference groups. Tungusic speakers from Northeast China carrying the C2b1b lineage and Mongolic-speaking populations from northern China with C2b1b and Q2a2b lineages also contributed to their distinct genomic characteristics. The origin center and phylogeographical distribution analyses indicate that these founding lineages originated from the Mongolian Plateau and surrounding areas. Ancient DNA evidence from the Mongolian Plateau and Amur River Basin demonstrates the genetic continuity of Neolithic Mongolian peoples, linking their migration directly to the spread of Trans-Eurasian languages [76]. Autosome-based evidence also suggested a recent genomic admixture between northern Chinese populations and western steppe herders, as well as local hunter-gatherer and Han-related farmer sources, which contributed to the genomic and linguistic diversity of Mongolians and their Han neighbors at the crossroads between Siberia and northern China [77]. This work observed population differentiation among Tibeto-Burman and Sinitic speakers, Sinitic and Tai-Kadai peoples, and northern and southern Han Chinese through PCA, MDS, and neighbor-joining clustering patterns. Admixture models based on qpGraph and coalescence theory reveal that deep Asian ancestry and Neolithic millet farmer ancestry contributed to the formation of highland Tibeto-Burman-speaking populations [78, 79]. Our findings on the frequency spectrum and hot spot patterns of key lineages indicate that D1a1b and its sublineages have the highest frequency among highland populations (Figs. 4 and 6), consistent with the long-term settlement of ancient millet farmers from northern China and biological adaptation to high-altitude environments on the Tibetan Plateau [80–82].

Genetic researchers identified genetic distinctions between northern and southern Han Chinese populations using high-density SNP genotyping data a decade ago [83]. Our study confirms this pattern through extensive paternal genomic resources and reconstructs the differentiated population history using multiple genetic models. Leveraging our integrated database, we provided a high-resolution view of differentiated lineages, shedding light on the founding lineages that contributed to the gene pool of northern and southern East Asians. Our findings indicate that the O2a2a and O2a1a lineages contributed to both northern and southern populations, while the O1a1a and O2a2b lineages played a primary role in southern East Asians, likely linked to the spread of rice and millet agriculture. Ancient genomes from the Yellow River Basin (Henan and Shandong provinces) and the Yangtze River Basin (Fujian, Guangxi, and Taiwan) suggest that genetic differentiation between northern and southern East Asians began in the early Neolithic period. Gene flow from Chinese farmers significantly influenced the genetic structure of ancient Southeast Asians, impacting the ancestors of Austronesian, Austroasiatic, Tai-Kadai, and Hmong-Mien peoples [84–86]. While our work highlights the deep paternal genetic history of Chinese populations, we acknowledge its limitations. We provided a large-scale panel of genotype data and the genetic diversity of linguistically diverse Chinese populations, along with the highest-resolution lineage frequency spectrum based on the CPGDD. However, additional high-coverage Y-chromosome sequencing data from underrepresented populations are needed to further elucidate the evolution of human Y-chromosome structures and paternal evolutionary history. Recent advances in technologies such as third-generation sequencing (TGS), the human pangenome project, and T2T genome assembly offer new opportunities to comprehensively understand human evolution and its relationship with Y-chromosome structures [13, 14, 87]. Ancient DNA offers direct insights into populations' evolutionary history, while the mutation trajectory of the paternally inherited Y-chromosome provides indirect inferences. Recently, genomic research and medical applications have primarily focused on Europeans and their descendants. This focus has hindered the transfer of genomic determinants of complex traits and evolutionary traces to East Asians due to differences in demographic history and genetic diversity. To address this gap, we developed a high-resolution panel with 2850 targeted genomic regions and 2999 phylogenetically informative SNPs, launching the YHC to explore the fine-scale paternal history of Chinese populations. The YHSeqY3000 panel offers the highest resolution for determining East Asian-specific Y-chromosome lineages, providing new avenues for forensic parentage identification, pedigree search, and paternal biogeographical ancestry inference. Our population genetic analyses, based on the integrated YHC database, identified linguistically diverse Chinese populations with distinct paternal gene pools. The Y-chromosome variation patterns partly aligned with genetic diversity patterns observed in Chinese populations through ancient and modern autosomal genome-wide variation-based admixture models. The correlation between founding lineages and their frequency distribution indicates significant genetic differentiation between northern and southern East Asians and a distinction between western and eastern East Asians. We provided a fine-scale lineage frequency spectrum and identified the geographical origins of key founding lineages based on the CPGDD. These findings highlighted significant differentiation between northern and southern and between western and eastern East Asians, explained by the observed genetic differences in paternal structure among ethnolinguistically diverse Chinese populations. The second phase of the YHC, utilizing high-coverage NGS and TGS techniques, could offer a comprehensive view of Y-chromosome genomic diversity, enhancing our CPGDD resolution from SNPs to STRs and other SVs, and advancing the understanding of Y-chromosome structure and human evolution.

Conclusion

This study comprehensively validated the sequencing performance and forensic and population genetic functions of the YHSeqY3000 panel via a well-designed strategy, which provides a high-resolution analysis of Y-chromosome genetic diversity in Chinese populations, uncovering fine-scale paternal substructures associated with geography, ethnicity, and evolutionary history. By developing the YHSeqY3000 panel and leveraging extensive genomic resources, we illuminate the genetic differentiation between northern and southern East Asians, the contributions of Neolithic agricultural expansions to the modern Chinese gene pool, and the influence of Western Eurasian ancestry in northwestern China. These findings highlight the value of large-scale genomic studies in addressing the underrepresentation of diverse populations, offering new insights into human evolution, migration, and admixture. Future advances in sequencing technologies and expanded datasets promise to further refine the understanding of Y-chromosome diversity and its implications for genetics, forensic science, and anthropology.

Methods

Marker selection and panel design

We selected Y-SNPs to create a population-specific NGS panel based on our internal integrated CPGDD database. The inclusion criteria for Y-SNPs were as follows:

(i) the frequency of the terminal haplogroup in Chinese populations was no more than 4‰; (ii) biallelic Y-SNPs without reverse mutation; (iii) Y-InDels were excluded; (iv) Y-SNPs with high heterozygosity in whole-genome/capture sequencing data in our database were excluded; (v) dominant Y-chromosomal lineages of Chinese populations with a TMRCA within 3000 years were included; (vi) primers had to be designable for NGS platforms. We extracted sequence information for the screened Y-SNPs from the NCBI database (https://www.ncbi.nlm.nih.gov/) using their rs numbers or physical locations. We employed the AIdesign tool (iGeneTech, Beijing, China) to design PCR and sequencing primers with the following specifications: (i) primers were designed as a single mixed pool; (ii) amplicon length was less than 150 bp; (iii) multiple primer pooling and sequencing experiments were conducted, with primers redesigned for Y-SNPs exhibiting low sequencing performance or replaced with parallel Y-SNPs from the same lineage to improve coverage rate (the proportion of target regions sequenced); (iv) the 5′ end of the primer could contain a non-target Y-SNP.

Sample preparation

We used control DNA 9948 (Promega, Madison, USA) to test repeatability, sensitivity, and stability. For repeatability testing, we amplified and sequenced three replicates of control DNA 9948 at the recommended input of 5 ng in three separate sequencing experiments. To assess sensitivity, we prepared a dilution series of control DNA 9948 at 5 ng, 2 ng, 1 ng, 0.5 ng, 0.1 ng, 50 pg, 10 pg, and 5 pg, and performed library preparation and sequencing in triplicate with the recommended number of PCR cycles. For low-concentration samples (0.5 ng, 0.1 ng, 50 pg, 10 pg, and 5 pg), we used a modified PCR protocol with three additional cycles to evaluate the effect of increased PCR cycles on sequencing quality. We also assessed the sequencing performance of the YHSeqY3000 panel on degraded DNA by digesting control DNA 9948 with DNase I (Thermo Fisher Scientific) for various durations (0, 1, 2, 4, 6, 8, and 15 min). To evaluate the stability of our newly developed panel, we added four common inhibitors—hemoglobin, indigo, humic acid, and EDTA—to the library preparation system at varying final concentrations: 50, 75, 100, and 125 μmol/L of hemoglobin; 6, 8, 10, and 12 mmol/L of indigo; 80, 100, 150, and 200 mg/L of humic acid; and 0.5, 0.8, 1.0, and 1.5 mmol/L of EDTA.

To assess the typing ability of the YHSeqY3000 panel on forensic case-type samples, we collected sixteen mock case-type DNA samples from blood cards, blood stains, cigarette butts, hair roots, and semen stains. For mixture testing, we mixed control DNA 9948 and 2800 M (Promega, Madison, USA) in various ratios (49:1, 29:1, 19:1, 9:1, 4:1, 1:1, 1:4, 1:9, 1:19, 1:29, and 1:49) to create different mixture samples. Library preparation followed standard protocols, and each mixture sample was sequenced twice to ensure reliable results. We also obtained DNA samples from duck, sheep, pig, dog, cat, rabbit, mouse, chicken, cow, *Streptococcus mutans* (*S. mutans*), *Candida albicans* (*C. albicans*), and *Escherichia coli* (*E. coli*) for species specificity testing. The YHSeqY3000 panel was designed to classify haplogroups of ethnolinguistically diverse Chinese populations on a fine scale. To evaluate its resolution and coverage, we collected peripheral blood samples from 2097 unrelated individuals across 41 geographically distinct populations representing 18 ethnically diverse groups. This included 92 Bai, 83 Bouyei, 54 Dai, 94 Dong, 33 Hani, 89 Hui, 87 Kazakh, 94 Korean, 36 Li, 85 Manchu, 91 Miao, 87 Mongolian, 93 Tibetan, 84 Tujia, 90 Yao, 89 Yi, 86 Zhuang, and 637 Han individuals. We used Y-chromosome data from the YHC phase 1 as reference populations in our population genetic analysis. The final merged dataset comprised 2375 individuals from 59 minority ethnic groups and 5661 Han Chinese individuals from 29 geographically diverse populations. We conducted genetic analyses on population sizes exceeding ten or thirty individuals, utilizing 26 Han Chinese populations and 32 minority ethnic groups with populations over 30 in the analysis.

DNA extraction, library preparation, and sequencing

We extracted DNA from human and non-human samples using the PureLink Genomic DNA Mini Kit (Thermo Fisher Scientific) and quantified it with the Qubit dsDNA HS Assay Kit (Thermo Fisher Scientific) on an Invitrogen Qubit 3.0 Fluorometer, following the manufacturer's instructions. We diluted the quantified DNA to 2.0 ng/μL and stored it at − 20 °C until library preparation. For library preparation, we used a customized primer pool and the MultipSeq Custom Panel (iGeneTech, Beijing, China) according to the manufacturer's protocol. The recommended DNA input amount per reaction was 5 ng, except for sensitivity testing. We amplified DNA samples and performed incubation steps using the ProFlex 96-Well PCR System (Thermo Fisher Scientific) with the following thermal conditions: lid temperature at 105 °C; initial denaturation at 95 °C for 3 min 30 s; 30 cycles of amplification at 98 °C for 20 s and 60 °C for 10 min; and a final extension at 72 °C for 5 min. We purified the PCR products using Agencourt AMPure XP beads (Beckman Coulter). The ligation reaction included 2.5 μL of Enhancer buffer M, 2 μL of UDI Index, 10 μL of IGT-EM808 polymerase mixture, 13.5 μL of purified PCR products, and 2 μL of ddH2O. The standard conditions for this reaction were as follows: incubation at 95 °C for 3 min 30 s; 9 cycles

of amplification at 98 °C for 20 s, 58 °C for 1 min, and 72 °C for 30 s; with a final hold at 72 °C for 5 min. Each DNA library was purified using Agencourt AMPure XP beads and quantified with an Invitrogen Qubit 3.0 Fluorometer. We assessed the length and purity of each purified library using the Qsep400 Bio-Fragment Analyzer (BiOptic, Taiwan, China). Finally, we conducted sequencing on an MGISEQ-2000 sequencer (MGI, Shenzhen, China) in paired-end 150 bp mode.

Sequencing data processing

We used Trimmomatic v.0.38 [88] to trim and crop raw FASTQ data and remove adapters. We aligned the raw sequencing reads to the human genome reference assembly GRCh37 using BWA-MEM v.0.7.12 [89]. We then sorted the aligned reads by chromosome and position using SAMtools v.1.9. For variant discovery, we employed the GATK HaplotypeCaller, CombineGVCFs, and GenotypeGVCFs modules [90]. We recommended a minimum sequencing depth of 30X per tested amplicon or target Y-SNP, a mapping quality greater than 20, a base quality above Q30, and a major allele proportion exceeding 90%. We evaluated the sequencing performance of the YHSeqY3000 panel based on average sequencing depth, the proportion of Y-SNPs exceeding 30X, the proportion of Y-SNPs reaching Q30, and the mapping, capture, and coverage rates. The mapping rate measures the ratio of reads mapped to the reference genome to the number of raw sequencing reads. The capture rate refers to the proportion of reads in the target regions compared to the number of reads mapped to the reference genome. The coverage rate indicates the ratio of the sequenced length of target regions to their preset length.

Haplogroup classification

We classified NRY haplogroups using in-house scripts based on our reconstructed phylogenetic tree. To ensure compatibility with previous studies that used the YCC nomenclature system, we assigned Y-chromosomal haplogroups using Y-LineageTracker [91] and HaploGrouper [92]. These assignments were based on the ISOGG Y-DNA Haplogroup Tree 2019–2020 (https://isogg.org/tree/index.html).

Estimation of statistical parameters

Using Arlequin v.3.5.1.3, we calculated haplotype and haplogroup frequencies and computed HD and HGD using the formula proposed by Nei and Tajima: HD/HGD $= n(1-\Sigma pi^2)/(n-1)$, where n is the total number of observed haplotypes or haplogroups, and pi is the frequency of the ith haplotype or haplogroup [93]. We determined the DP as the ratio between the number of observed haplotypes and the total number of haplotypes. MP was estimated using the formula: MP $= \Sigma pi^2$.

Inference of population structure

To characterize the paternal genetic landscape of ethnolinguistically diverse Chinese populations, we integrated publicly available haplotype and haplogroup data from the YHC genomic resource [94, 95] into this study. We constructed a maximum likelihood phylogenetic tree using RAxML (Randomized Axelerated Maximum Likelihood) v.8.2.12 [96] and visualized relationships between individual genotypes with haplotype networks generated by PopART [97]. To analyze the genetic structure of Chinese populations at the group and language family levels, we conducted PCA and MDS analysis using Y-LineageTracker [91], based on fourth-level haplogroup frequencies. We estimated pairwise genetic distances (Fst) with Y-LineageTracker and constructed a neighbor-joining tree using MEGA 7 [98]. Correlation analysis between haplogroup frequencies and latitude/longitude or Fst genetic distances was performed with the R package to identify haplogroups contributing to genetic differentiation among geographically distinct populations. Spatial autocorrelation analysis of haplogroup distribution patterns was performed using ArcGIS.

Abbreviations

T2T	Telomere-To-Telomere
SNPs	Single nucleotide polymorphisms
SVs	Structural variations
RFLPs	Restriction fragment length polymorphisms
InDels	Insertion/deletion
STRs	Short tandem repeats
YCC	Y Chromosome Consortium
ISOGG	International Society of Genetic Genealogy
CPGDD	Chinese Paternal Genomic Diversity Database
10K_CPGDP	10K Chinese People Genomic Diversity Project
YHC	YanHuang cohort
HD	Haplotype diversity
HGD	Haplogroup diversity
DP	Discrimination probability
MP	Match probability

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12915-025-02122-0.

Additional file 1: Supplementary Table S1. The detailed information on newly sequenced Chinese populations from the YanHuang cohort

Additional file 2: Supplementary Table S2. Detailed information on Y-SNPs included in the newly developed YHSeqY3000 panel

Additional file 3: Supplementary Table S3. The results of sequencing performance of the newly developed YHSeqY3000 panel

Additional file 4: Supplementary Table S4. The assigned haplogroups of 2097 newly genotyped individuals from 41 ethnolinguistically diverse Chinese populations

Additional file 5: Supplementary Fig.ure S1. Principal component analysis (PCA) showed the genetic relationships among 88 Chinese populations

Acknowledgements

We thank all the volunteers who participated in this project.
Full author lists of 10K_CPGDP consortium

Chao Liu[1], Guanglin He[2], Mengge Wang[2], Renkuan Tang[3], Libing Yun[4], Junbao Yang[5], Chuan-Chao Wang[6], Jiangwei Yan[7], Bofeng Zhu[8], Liping Hu[9], Shengjie Nie[9], Hongbing Yao[10]
[1]Anti-Drug Technology Center of Guangdong Province, Guangzhou, 510220, China
[2]Institute of Rare Diseases, West China Hospital of Sichuan University, Sichuan University, Chengdu, 610000, China
[3]Department of Forensic Medicine, College of Basic Medicine, Chongqing Medical University, Chongqing, 400331, China
[4]West China School of Basic Science and Forensic Medicine, Sichuan University, Chengdu, 610041, China
[5]School of Basic Medicine and Forensic Medicine, North Sichuan Medical College, Nanchong, Sichuan, 637007, China
[6]State Key Laboratory of Cellular Stress Biology, School of Life Sciences, Xiamen University, Xiamen, 361005, China
[7]School of Forensic Medicine, Shanxi Medical University, Jinzhong, 030001, China
[8]Guangzhou Key Laboratory of Forensic Multi-Omics for Precision Identification, School of Forensic Medicine, Southern Medical University, Guangzhou, 510220, China
[9]School of Forensic Medicine, Kunming Medical University, Kunming, 650500, China
[10]Belt and Road Research Center for Forensic Molecular Anthropology, Gansu University of Political Science and Law, Lanzhou, 730000, China

Code availability
No custom code has been used in this work.

Authors' contributions
G.H., M.W., and C.L. conceived and supervised the project. G.H. and M.W. collected the samples. Z.W., Y.L., G.H., and M.W. extracted the genomic DNA and performed the genome sequencing. Z.W., G.H., M.W., and K.L. did variant calling. Z.W., S.D., Y.H.L., Y.L., X.L., L.W., H.Y., B.L., L.H., S.N., K.Z., C.L., Q.S., and G.H. performed population genetic analysis. H.Y., B.L., L.H., S.N., G.H., and M.W. drafted the manuscript. S.D., G.H., Y.M., Q.S., Y.M., M.W., and C.L. revised the manuscript. All authors read and approved the final manuscript.

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

Data availability
All data generated or analyzed during this study are included in this published article and its supplementary information files. We followed the regulations of the Ministry of Science and Technology of the People's Republic of China. The raw data were submitted to the Zenodo database (https://zenodo.org/records/10591488).

Declarations

Ethics approval and consent to participate
The Medical Ethics Committee of West China Hospital of Sichuan University approved this study (2023–1321). This study was conducted in accordance with the principles of the Helsinki Declaration.

Consent for publication
Not applicable.

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

Author details
[1]Institute of Rare Diseases, Frontiers Science Center for Disease-Related Molecular Network, West China Hospital, Sichuan University, Chengdu 610000, Sichuan, China. [2]Center for Archaeological Science, Sichuan University, Chengdu 610000, China. [3]School of Basic Medical Sciences, North Sichuan Medical College, Nanchong 637100, China. [4]School of International Tourism and Culture, Guizhou Normal University, Guiyang 550025, China. [5]MoFang Human Genome Research Institute, Tianfu Software Park, Chengdu 610042, Sichuan, China. [6]School of Ethnology and Anthropology, Inner Mongolia Normal University, Hohhot 010028, Inner Mongolia, China. [7]School of Forensic Medicine, Kunming Medical University, Kunming 650500, China. [8]Department of Forensic Medicine, College of Basic Medicine, Chongqing Medical University, Chongqing 400331, China. [9]Anti-Drug Technology Center of Guangdong Province, Guangzhou 510230, China. [10]Guangzhou Key Laboratory of Forensic Multi-Omics for Precision Identification, School of Forensic Medicine, Southern Medical University, Guangzhou 510515, China. [11]Department of Medical Genetics, Frontiers Science Center for Disease-Related Molecular Network, West China Hospital, Sichuan University, Chengdu 610041, Sichuan, China. [12]Institute of Forensic Science, Ministry of Public Security, Beijing 100038, China. [13]Department of Oto-Rhino-Laryngology, West China Hospital of Sichuan University, Chengdu 610000, China.

Received: 1 February 2024 Accepted: 7 January 2025
Published online: 21 January 2025

References
1. 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):3426–3440 e3419.
2. He G, Wang M, Luo L, Sun Q, Yuan H, Lv H, Feng Y, Liu X, Cheng J, Bu F, et al. Population genomics of Central Asian peoples unveil ancient Trans-Eurasian genetic admixture and cultural exchanges. Life. 2024;2(11):554–62.
3. Sun Y, Wang M, Sun Q, Liu Y, Duan S, Wang Z, Zhou Y, Zhong J, Huang Y, Huang X, et al. Distinguished biological adaptation architecture aggravated population differentiation of Tibeto-Burman-speaking people. J Genet Genomics. 2024;51(5):517–30.
4. Li X, Wang M, Su H, Duan S, Sun Y, Chen H, Wang Z, Sun Q, Yang Q, Chen J, et al. Evolutionary history and biological adaptation of Han Chinese people on the Mongolian Plateau. Life. 2024;2(6):296–313.
5. He G, Wang M, Luo L, Sun Q, Yuan H, Lv H, Feng Y, Liu X, Cheng J, Bu F, et al. Population genomics of Central Asian peoples unveil ancient Trans-Eurasian genetic admixture and cultural exchanges. hLife. 2024;2(11):554–62.
6. Wang M, Huang Y, Liu K, Wang Z, Zhang M, Yuan H, Duan S, Wei L, Yao H, Sun Q, et al. Multiple human population movements and cultural dispersal events shaped the landscape of Chinese paternal heritage. Mol Biol Evol 2024, 41(7):2023.2008.2028.555114.
7. Wang M, Chen H, Luo L, Huang Y, Duan S, Yuan H, Tang R, Liu C, He G. Forensic investigative genetic genealogy: expanding pedigree tracing and genetic inquiry in the genomic era. J Genet Genomics. 2024.
8. Nurk S, Koren S, Rhie A, Rautiainen M, Bzikadze AV, Mikheenko A, Vollger MR, Altemose N, Uralsky L, Gershman A, et al. The complete sequence of a human genome. Science. 2022;376(6588):44–53.
9. Hoyt SJ, Storer JM, Hartley GA, Grady PGS, Gershman A, de Lima LG, Limouse C, Halabian R, Wojenski L, Rodriguez M, et al. From telomere to telomere: the transcriptional and epigenetic state of human repeat elements. Science. 2022;376(6588):eabk3112.
10. Gershman A, Sauria MEG, Guitart X, Vollger MR, Hook PW, Hoyt SJ, Jain M, Shumate A, Razaghi R, Koren S, et al. Epigenetic patterns in a complete human genome. Science. 2022;376(6588):eabj5089.
11. Altemose N, Logsdon GA, Bzikadze AV, Sidhwani P, Langley SA, Caldas GV, Hoyt SJ, Uralsky L, Ryabov FD, Shew CJ, et al. Complete genomic and epigenetic maps of human centromeres. Science. 2022;376(6588):eabl4178.
12. Aganezov S, Yan SM, Soto DC, Kirsche M, Zarate S, Avdeyev P, Taylor DJ, Shafin K, Shumate A, Xiao C, et al. A complete reference genome improves analysis of human genetic variation. Science. 2022;376(6588):eabl3533.
13. Rhie A, Nurk S, Cechova M, Hoyt SJ, Taylor DJ, Altemose N, Hook PW, Koren S, Rautiainen M, Alexandrov IA, et al. The complete sequence of a human Y chromosome. Nature. 2023;621(7978):344–54.
14. Hallast P, Ebert P, Loftus M, Yilmaz F, Audano PA, Logsdon GA, Bonder MJ, Zhou W, Hops W, Kim K, et al. Assembly of 43 human Y

chromosomes reveals extensive complexity and variation. Nature. 2023;621(7978):355–64.
15. Wang M, Sun Q, Feng Y, Wei LH, Liu K, Luo L, Huang Y, Zhou K, Yuan H, Lv H, et al. Paleolithic divergence and multiple Neolithic expansions of ancestral nomadic emperor-related paternal lineages. J Genet Genomics. 2024.
16. Wang M, Liu Y, Luo L, Feng Y, Wang Z, Yang T, Yuan H, Liu C, He G. Genomic insights into Neolithic founding paternal lineages around the Qinghai-Xizang plateau using integrated YanHuang resource. Science. 2024;27:111456.
17. Wang M, Huang Y, Liu K, Wang Z, Zhang M, Yuan H, Duan S, Wei L, Yao H, Sun Q, et al. Multiple human population movements and cultural dispersal events shaped the landscape of Chinese paternal heritage. Mol Biol Evol. 2024;41(7):msae122.
18. Hammer MF. A recent common ancestry for human Y chromosomes. Nature. 1995;378(6555):376–8.
19. Oota H, Settheetham-Ishida W, Tiwawech D, Ishida T, Stoneking M. Human mtDNA and Y-chromosome variation is correlated with matrilocal versus patrilocal residence. Nat Genet. 2001;29(1):20–1.
20. Gunnarsdottir ED, Nandineni MR, Li M, Myles S, Gil D, Pakendorf B, Stoneking M. Larger mitochondrial DNA than Y-chromosome differences between matrilocal and patrilocal groups from Sumatra. Nat Commun. 2011;2:228.
21. Ellis N, Hammer M, Hurles ME, Jobling MA, Karafet T, King TE, de Knijff P, Pandya A, Redd A, Santos FR, et al. A nomenclature system for the tree of human Y-chromosomal binary haplogroups. Genome Res. 2002;12(2):339–48.
22. 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.
23. 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.
24. Wen B, Xie X, Gao S, Li H, Shi H, Song X, Qian T, Xiao C, Jin J, Su B, et al. Analyses of genetic structure of Tibeto-Burman populations reveals sex-biased admixture in southern Tibeto-Burmans. Am J Hum Genet. 2004;74(5):856–65.
25. Shi H, Dong YL, Wen B, Xiao CJ, Underhill PA, Shen PD, Chakraborty R, Jin L, Su B. Y-chromosome evidence of southern origin of the East Asian-specific haplogroup O3–M122. Am J Hum Genet. 2005;77(3):408–19.
26. Zhong H, Shi H, Qi XB, Duan ZY, Tan PP, Jin L, Su B, Ma RZ. Extended Y chromosome investigation suggests postglacial migrations of modern humans into East Asia via the northern route. Mol Biol Evol. 2011;28(1):717–27.
27. Qi X, Cui C, Peng Y, Zhang X, Yang Z, Zhong H, Zhang H, Xiang K, Cao X, Wang Y, et al. Genetic evidence of paleolithic colonization and neolithic expansion of modern humans on the tibetan plateau. Mol Biol Evol. 2013;30(8):1761–78.
28. Peng MS, Zhang YP. Sex-biased adaptation shapes uniparental gene pools in Tibetans. Sci China Life Sci. 2024;67(3):611–3.
29. Jobling MA, Tyler-Smith C. Human Y-chromosome variation in the genome-sequencing era. Nat Rev Genet. 2017;18(8):485–97.
30. Kayser M. Forensic use of Y-chromosome DNA: a general overview. Hum Genet. 2017;136(5):621–35.
31. Wang Z, Wang M, Liu K, Yuan H, Duan S, Liu Y, Luo L, Jiang X, Chen S, Wei L, et al. Paternal genomic resources from the YanHuang cohort suggested a weakly-differentiated multi-source admixture model for the formation of Han's founding ancestral lineages. Genomics, Proteomics & Bioinformatics. 2023:2023.2011.2008.566335.
32. van Oven M, van den Tempel N, Kayser M. A multiplex SNP assay for the dissection of human Y-chromosome haplogroup O representing the major paternal lineage in East and Southeast Asia. J Hum Genet. 2012;57(1):65–9.
33. Yin C, Ren Y, Adnan A, Tian J, Guo K, Xia M, He Z, Zhai D, Chen X, Wang L, et al. Title: developmental validation of Y-SNP pedigree tagging system: a panel via quick ARMS PCR. Forensic Sci Int Genet. 2020;46:102271.
34. 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.
35. 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.
36. 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.
37. 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.
38. 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.
39. 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 Science International: Genetics Supplement Series. 2017;6:e96–8.
40. 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.
41. Liu J, Jiang L, Zhao M, Du W, Wen Y, Li S, Zhang S, Fang F, Shen J, He G, et al. Development and validation of a custom panel including 256 Y-SNPs for Chinese Y-chromosomal haplogroups dissection. Forensic Sci Int Genet. 2022;61:102786.
42. Tao R, Li M, Chai S, Xia R, Qu Y, Yuan C, Yang G, Dong X, Bian Y, Zhang S, et al. Developmental validation of a 381 Y-chromosome SNP panel for haplogroup analysis in the Chinese populations. Forensic Sci Int Genet. 2023;62:102803.
43. He G, Wang M, Miao L, Chen J, Zhao J, Sun Q, Duan S, Wang Z, Xu X, Sun Y, et al. 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. 2023;17(1):29.
44. Luo L, Chao L, Mengge W, Yunhui L, Jianbo L, Fengxiao B, Hunjun Y, Renkuan T, Guanglin H. Sequencing and characterizing human mitochondrial genomes in the biobank-based genomic research paradigm. SCIENCE CHINA Life Sciences. 2024.
45. He G, Yao H, Duan S, Luo L, Sun Q, Tang R, Chen J, Wang Z, Sun Y, Li X, et al. Pilot work of the 10K Chinese People Genomic Diversity Project along the Silk Road suggests a complex east–west admixture landscape and biological adaptations. SCIENCE CHINA Life Sciences. 2024.
46. 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-+.
47. Mao XW, Zhang HC, Qiao SY, Liu YC, Chang FQ, Xie P, Zhang M, Wang TY, Li MA, Cao P, et al. The deep population history of northern East Asia from the Late Pleistocene to the Holocene. Cell. 2021;184(12):3256-+.
48. Wang TY, Wang W, Xie GM, Li Z, Fan XC, Yang QP, Wu XC, Cao P, Liu YC, Yang RW, et al. Human population history at the crossroads of East and Southeast Asia since 11,000 years ago. Cell. 2021;184(14):3829-+.
49. Tao L, Yuan H, Zhu K, Liu X, Guo J, Min R, He H, Cao D, Yang X, Zhou Z, et al. Ancient genomes reveal millet farming-related demic diffusion from the Yellow River into southwest China. Curr Biol. 2023;33(22):4995–5002 e4997.
50. Robbeets M, Bouckaert R, Conte M, Savelyev A, Li T, An DI, Shinoda KI, Cui Y, Kawashima T, Kim G, et al. Triangulation supports agricultural spread of the Transeurasian languages. Nature. 2021;599(7886):616–21.
51. Yang MA, Fan XC, Sun B, Chen CY, Lang JF, Ko YC, Tsang CH, Chiu HL, Wang TY, Bao QC, et al. Ancient DNA indicates human population shifts and admixture in northern and southern China. Science. 2020;369(6501):282-+.
52. Xiong J, Wang R, Chen G, Yang Y, Du P, Meng H, Ma M, Allen E, Tao L, Wang H, et al. Inferring the demographic history of Hexi Corridor over the past two millennia from ancient genomes. Science Bulletin. 2024;69(5):606–11.

53. Zhang F, Ning C, Scott A, Fu Q, Bjorn R, Li W, Wei D, Wang W, Fan L, Abuduresule I, et al. The genomic origins of the Bronze Age Tarim Basin mummies. Nature. 2021;599(7884):256–61.
54. Kumar V, Wang W, Zhang J, Wang Y, Ruan Q, Yu J, Wu X, Hu X, Wu X, Guo W, et al. Bronze and Iron Age population movements underlie Xinjiang population history. Science. 2022;376(6588):62–9.
55. Wang M, Yao H, Sun Q, Duan S, Tang R, Chen J, Wang Z, Sun Y, Li X, Wang S, et al. Pilot work of the 10K Chinese People Genomic Diversity Project along the Silk Road suggests a complex east–west admixture landscape and biological adaptations. Science China-life Sciences. 2024:2023.2002.2026.530053.
56. Luo L, Wang M, Liu Y, Li J, Bu F, Yuan H, Tang R, Liu C, He G. Sequencing and characterizing human mitochondrial genomes in the biobank-based genomic research paradigm. Science China-Life Sciences. 2024.
57. Skaletsky H, Kuroda-Kawaguchi T, Minx PJ, Cordum HS, Hillier L, Brown LG, Repping S, Pyntikova T, Ali J, Bieri T, et al. The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes. Nature. 2003;423(6942):825–37.
58. Novembre J, Johnson T, Bryc K, Kutalik Z, Boyko AR, Auton A, Indap A, King KS, Bergmann S, Nelson MR, et al. Genes mirror geography within Europe. Nature. 2008;456(7218):98–101.
59. Liu YC, Mao XW, Krause J, Fu QM. Insights into human history from the first decade of ancient human genomics. Science. 2021;373(6562):1479–84.
60. 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):msac045.
61. 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.
62. 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.
63. 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.
64. 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):149–157 e143.
65. 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.
66. Wei LH, Yan S, Lu Y, Wen SQ, Huang YZ, Wang LX, Li SL, Yang YJ, Wang XF, Zhang C, et al. Whole-sequence analysis indicates that the Y chromosome C2*-Star Cluster traces back to ordinary Mongols, rather than Genghis Khan. Eur J Hum Genet. 2018;26(2):230–7.
67. Cheng S, Xu Z, Bian SZ, Chen X, Shi YF, Li YR, Duan YY, Liu Y, Lin JX, Jiang Y, et al. The STROMICS genome study: deep whole-genome sequencing and analysis of 10K Chinese patients with ischemic stroke reveal complex genetic and phenotypic interplay. Cell Discovery. 2023;9(1):75.
68. Wang C, Dai J, Qin N, Fan J, Ma H, Chen C, An M, Zhang J, Yan C, Gu Y, et al. Analyses of rare predisposing variants of lung cancer in 6,004 whole genomes in Chinese. Cancer Cell. 2022;40(10):1223–1239 e1226.
69. Cong PK, Bai WY, Li JC, Yang MY, Khederzadeh S, Gai SR, Li N, Liu YH, Yu SH, Zhao WW, et al. Genomic analyses of 10,376 individuals in the Westlake BioBank for Chinese (WBBC) pilot project. Nat Commun. 2022;13(1):2939.
70. 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.
71. Jiang T, Guo H, Liu Y, Li G, Cui Z, Cui X, Liu Y, Li Y, Zhang A, Cao S, et al. A comprehensive genetic variant reference for the Chinese population. Sci Bull (Beijing). 2024;69(24):3820–25.
72. 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.
73. Zhao GB, Miao L, Wang M, Yuan JH, Wei LH, Feng YS, Zhao J, Kang KL, Zhang C, Ji AQ, et al. Developmental validation of a high-resolution panel genotyping 639 Y-chromosome SNP and InDel markers and its evolutionary features in Chinese populations. BMC Genomics. 2023;24(1):611.
74. Ning Z, Tan X, Yuan Y, Huang K, Pan Y, Tian L, Lu Y, Wang X, Qi R, Lu D, et al. Expression profiles of east-west highly differentiated genes in Uyghur genomes. Natl Sci Rev. 2023;10(4):nwad077.
75. Pan Y, Zhang C, Lu Y, Ning Z, Lu D, Gao Y, Zhao X, Yang Y, Guan Y, Mamatyusupu D, et al. Genomic diversity and post-admixture adaptation in the Uyghurs. Natl Sci Rev. 2022;9(3):nwab124.
76. 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.
77. 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.
78. Lu D, Lou H, Yuan K, Wang X, Wang Y, Zhang C, Lu Y, Yang X, Deng L, Zhou Y, et al. Ancestral origins and genetic history of Tibetan highlanders. Am J Hum Genet. 2016;99(3):580–94.
79. 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.
80. Bai F, Liu Y, Wangdue S, Wang T, He W, Xi L, Tsho Y, Tsering T, Cao P, Dai Q, et al. Ancient genomes revealed the complex human interactions of the ancient western Tibetans. Curr Biol. 2024;34(12):2594–2605.e2597.
81. Zheng WS, He YX, Guo YB, Yue T, Zhang H, Li J, Zhou B, Zeng XR, Li LY, Wang B, et al. Large-scale genome sequencing redefines the genetic footprints of high-altitude adaptation in Tibetans. Genome Biol. 2023;24(1):73.
82. Wang HR, Yang MA, Wangdue S, Lu HL, Chen HH, Li LH, Dong GH, Tsring T, Yuan HB, He W, et al. Human genetic history on the Tibetan Plateau in the past 5100 years. Science advances. 2023;9(11):eadd5582.
83. Chen J, Zheng H, Bei JX, Sun L, Jia WH, Li T, Zhang F, Seielstad M, Zeng YX, Zhang X, et al. Genetic structure of the Han Chinese population revealed by genome-wide SNP variation. Am J Hum Genet. 2009;85(6):775–85.
84. 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.
85. 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.
86. 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.
87. Zhou Y, Zhan X, Jin J, Zhou L, Bergman J, Li X, Rousselle MMC, Belles MR, Zhao L, Fang M, et al. Eighty million years of rapid evolution of the primate Y chromosome. Nat Ecol Evol. 2023;7(7):1114–30.
88. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30(15):2114–20.
89. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25(14):1754–60.
90. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010;20(9):1297–303.
91. Chen H, Lu Y, Lu D, Xu S. Y-LineageTracker: a high-throughput analysis framework for Y-chromosomal next-generation sequencing data. BMC Bioinformatics. 2021;22(1):114.
92. 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.
93. Excoffier L, Lischer HE. Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. Mol Ecol Resour. 2010;10(3):564–7.

94. Wang Z, Wang M, Liu K, Yuan H, Duan S, Liu Y, Luo L, Jiang X, Chen S, Wei L. Paternal genomic resources from the YanHuang cohort suggested a weakly-differentiated multi-source admixture model for the formation of Han's founding ancestral lineages. bioRxiv. 2023:2023.2011. 2008.566335.
95. Wang M, Huang Y, Liu K, Yuan H, Duan S, Wang Z, Wei L, Yao H, Sun Q, Zhong J. Ancient farmer and steppe pastoralist-related founding lineages contributed to the complex landscape of episodes in the diversification of Chinese paternal lineages. bioRxiv. 2023;2023.2008. 2028.555114.
96. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 2014;30(9):1312–3.
97. Leigh JW, Bryant D, Nakagawa S. popart: full-feature software for haplotype network construction. Methods Ecol Evol. 2015;6(9):1110–6.
98. Kumar S, Stecher G, Tamura K. MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. Mol Biol Evol. 2016;33(7):1870–4.

Publisher's Note