Full text open access online (Since 2001)

PRINT: ISSN 0972-3757 ONLINE: ISSN 2456-6330
Int J Hum Genet, 24(4): 332-344 (2024)
DOI: 10.31901/24566322.2024/24.04.886

Whole-Y Sequencing Provided Clearer O Haplogroup Subclade Determinations Than Traditional Techniques Among Male Hans in Henan (Central China)

Hao Wang[1,2,#], Lu Yu[1,3,#], Ning Wang[1,#], Ruixue Peng[1,4], Zhongqian Guo[1,5], Lulu Wang[1], Lanhai Wei[6,7,*], Hong'en Xu[8,*] and Zhaoshu Zeng[1,*]

[1]Department of Forensic Medicine, School of Basic Medical Sciences, Zhengzhou University, Zhengzhou 450001, Henan, China
[2] Material Evidence Identification Center, Public Security Bureau of Xizang Autonomous Region, Lasa 850000, Xizang Autonomous Region, China
[3]Judicial Expertise Center, The 2nd People's Hospital of Zhengzhou City, Zhengzhou 450052, Henan, China
[4]Department of Forensic Sciences, Public Security Bureau of Pingdingshan City, Pingdingshan 467036, Henan, China
[5]Judicial Expertise Center, Henan No.3 Provincial People's Hospital, Zhengzhou 450006, Henan, China
[6]School of Ethnology and Anthropology, Inner Mongolia Normal University, Hohhot 010028, Inner Mongolia Autonomous Region, China
[7]Department of Anthropology and Ethnology, Institute of Anthropology, Xiamen University, Xiamen 361005, Fujian, China
[8]Precision Medicine Center, Academy of Medical Science, Zhengzhou University, Zhengzhou 450052, Henan, China

KEYWORDS DNA Endonucleolytic Cleavage. High-Throughput Nucleotide Sequencing. Anchored PCR. SNaPshot Minisequencing. Phylogenetic tree. Human Y chromosome.

ABSTRACT Binary polymorphisms in the non-recombinant portion of Y chromosome are useful on many occasions and have been investigated using traditional techniques, such as enzyme cleavage, sequence-specific PCR, and SNaPshot. However, Y haplogroup reports may not be sufficiently accurate compared to next-generation sequencing (NGS) data. Here, the researchers first analysed 205 samples from male Henan Han people using traditional methods and found 75.11 percent O and 11.71 percent C. Subsequently, the researchers analysed 55 samples via NGS. A comparison of each haplogroup of the same sample by traditional methods and by NGS showed consistency in main clade determination. However, in O subclade determinations, only one sample was identical between the two techniques (1/55=1.82%), two samples were absolutely different in subclades (2/55 = 3.64%), and the remaining 52 samples yielded clearer subclades by NGS/Yleaf than by traditional methods (52/55 = 94.54%). Overall, the study suggested that NGS/Yleaf was much clearer in Y haplogroup O subclade determinations.

INTRODUCTION

Binary polymorphisms in the non-recombinant part of the Y chromosome serve as valuable tools for tracking the source and movement patterns of contemporary human paternal lineages. As shown by the most recent Y haplogroup tree (ISOGG 2020), the topology has been clearly described, and the Y chromosome phylogeny of global males has been indicated.

There is a general consensus that Y haplogroup O is dominant in Chinese Han populations. In 2005, Shi et al. reported that O3-M122 had an average frequency of 44.3 percent, with O3-M122 in 2005 being shifted to O2-M122 in 2019 (Shi et al. 2005; ISOGG 2020). In addition to O, other haplogroups have been reported among Chinese Hans. For example, C-M130 was reported by Zhong and colleagues from Southeast Asian and East Asian populations (Zhong 2010; Zhong et

[*] *Addresses for correspondence:*
Lanhai Wei,
Professor
E-mail: Ryan.lh.wei@gmail.com
Dr. Hongen Xu,
E-mail: hongen_xu@zzu.edu.cn
Zhaoshu Zeng
Professor
E-mail: 1626660528@qq.com.

al. 2011). The haplogroups D and N as well as O and C were reported to account for 93 percent of the East Asian males according to a study in 2013 (Wang and Li 2013). With 27 Y short tandem repeat (Y-STR) and 143 Y single nucleotide polymorphism (Y-SNP) typing results from 3333 male Chinese individuals, Li reported in 2023 that O-M175, C-M130, R-M207, N-M231, and D-M174 are the five dominant haplogroups in China (Li et al. 2023). Most recently, He et al. reported a study on 1033 male Chinese individuals with a lately developed 639-plex Y-SNP panel, and O1a1, O1b1, O2a1, O2a2, O2b1, C2a1, C2b1 and Q1a1 were identified as the dominant Y-chromosomal lineages among Han Chinese individuals (He et al. 2023). Yu and Li also reported the historical or constant existence of O2a-M324 among Hans since Yangshao culture (5000-3300 B.C.E.) in northwestern China, and they found that O2a-M324, one of the most crucial non-recombining Y (NRY) lineages among Han Chinese people, existed in all samples of Qinghai males from the Han Dynasty (206 B.C.E.-220 C.E.) (Yu and Li 2021).

One recent study focused on male Henan Han individuals. In 2019, Lang et al. reported a phylogenetic analysis of male Henan Han individuals based on 27 Y-STRs and 143 Y-SNPs. They revealed that O, C, D, J, N, Q and R all existed in Henan Hans, with the haplogroup O-M175 accounting for 67.34 percent among the 199 male samples (Lang et al. 2019). Henan, literally meaning "the south bank of the Yellow River" in Chinese, along with the present-day Shaanxi and Shanxi provinces, is regarded as the cradle of Chinese civilization (Chen et al. 2009; Wang et al. 2021). The ancient Huaxia population, ancestors of the modern Han Chinese, primarily resided in these areas. The Yangshao culture, which thrived along the Yellow River and the Wei River valleys in northern China during the Neolithic period, is recognized as one of the earliest settled cultures in the region (Yangshao Culture 2020). The culture derived its name from Yangshao village in Henan Province, where the first excavations took place in 1921. Having lived in this area for more than 7000-8000 years, individuals of the ancient Huaxia population interacted with other contemporary indigenous populations for several thousand years (Yu and Li 2021; Chen et al. 2019), and these populations finally declared themselves as 'Han' with the name of the once poweful Han Dynasty (260 B.C.E.-220 C.E.) (Twitchett and Loewe 1986; Liu 2005; Poznik et al. 2016; Lu et al. 2020). Therefore, revealing the Y chromosomal haplogroup structures of male Hans in Henan Province is highly important. Although many other studies have analysed hundreds of thousands of Han individual samples, no additional recent Y haplogroup studies can be found especially on male Henan Han individuals.

Most of the previous studies included genotyping of Y-SNPs via traditional methods (Karafet et al. 2008), including DNA endonucleolytic cleavage (or enzyme digestion), anchored PCR (or sequence-specific PCR, SSP) technique, or SNaPshot minisequencing. The question of whether the data yielded by traditional methods are sufficiently reliable when compared to that of the high-throughput nucleotide sequencing (or the next-generation sequencing, NGS) results has subsequently risen from these studies.

NGS has become increasingly popular because of its sophisticated technical approaches, extremely detailed DNA reports and exponentially decreasing cost. Moreover, NGS has provided significant benefits to anthropology researchers. The variations provided by NGS are clear, the stability of haplogroup patterns is ensured, and the paradigm is finer than that obtained using traditional (or low-resolution) methods. Therefore, obtaining Y haplogroup data derived from NGS has posed a challenge to obtaining data published with traditional methods. The consistency or discrepancy of Y haplogroup determinations with these two techniques on the same sample should be compared and explored.

However, the Y haplogroup pattern of Henan Han individuals has not been thoroughly studied due to insufficient sample size or unavailability of enough Y-DNA markers (Lu et al. 2020; Lang et al. 2019). In this study, the researchers investigated 205 samples from male Henan Han people using traditional methods such as enzyme digestion, SSP, or SNaPshot, and then analysed 55 samples via high-throughput nucleotide sequencing on the Y chromosome to determine whether discrepancies in Y haplogroups existed between these two types of techniques. Averagely, the Y chromosome sequencing acquired

at least 57,600 Y-SNPs from each Y chromosome, as provided enough Y-SNPs for Yleaf 2.0 (Ralf et al. 2018) to precisely determine the haplogroup attribution of each male sample.

Objectives

The objectives of this study were as follows:

1. To compare the consistencies and discrepancies in terms of Y haplogroup of the same sample after 55 Y haplogroups were determined by traditional methods (enzyme digestion, sequence-specific PCRs, and SNaPshot) and then by NGS/Yleaf.
2. To obtain important information on the haplogroup pattern of modern male Henan Han individuals with enough samples.

MATERIALS AND METHODS

Sample Collection and DNA Extraction

205 blood samples were gathered from male Han individuals residing in Henan Province, China. All donors were unrelated and healthy, and came from families living in Henan Province for more than three generations. All the samples were subjected to enzyme digestion, SSP, or SNaPshot analysis. Among them, 55 samples were subjected to whole Y sequencing. All participants provided informed consent. In 2018, the use of all the samples was approved by Zhengzhou University's Ethics Committee with the ethical approval number ZZUGZR2018-3006, and in 2019, the use of some of these samples was approved by Xiamen University's Ethics Committee for Biological Research with ethical approval number XDYX2019009.

A DP-318 Kit (Tiangen Biotechnology, Beijing, China) and a QIAamp DNA Mini Kit (Qiagen, Dusseldorf, Germany) were both applied to extract genomic DNA from blood according to the manufacturers' manual. Obtained DNA was quantified with a NanoDrop 2000c (Thermo Fisher Scientific, Foster City, CA, USA). Finally, DNA were stored at -20 °C ready for PCR amplification.

Y-STR Typing with Capillary Electrophoresis

Twenty-nine Y-STR markers were amplified using the Microreader™ 29Y ID System (Microread Genetics Inc., Suzhou, Jiangsu, China). The 29 Y-STRs were DYS576, DYS389I, DYS635, DYS389II, DYS627, DYS460, DYS458, DYS19, GATA-H4, DYS448, DYS391, DYS456, DYS390, DYS438, DYS392, DYS518, DYS570, DYS437, DYS385a, DYS385b, DYS449, DYS393, DYS439, DYS481, DYF387S1a, DYF387S1b, DYS533, DYS549 and DYS643. A ProFlex 96-well PCR System (Thermo Fisher Scientific) was used for PCR according to the manufacturer's manual. Applied Biosystems 3130 and/or 3730 Genetic Analyzers (Applied Biosystems, Foster City, CA, USA) were used to electrophorese and detect the PCR products. GeneScan v.3.7 and Genotyper v.3.7 (Applied Biosystems) were used to analyse the electrophoresis results. Arlequin (version 3.5.2.2) (Excoffier and Lischer 2010) was used to analyse the haplotypes.

Y-SNP Marker Selection for Enzyme Digestion, SSP or SNaPshot

Eighteen Y-SNPs were selected for traditional low-resolution analysis, as shown in Table 1 (Yu 2018). Briefly, SNPs were selected from ISOGG (https://www.isogg.org), YHRD (https://yhrd.org, release 60), and the 1000 Genomes Project (https://www.ncbi.nlm.nih.gov/variation/tools/1000genomes/). The criteria for selecting Y-SNPs for the study were that the validated SNPs should provide clear branch information on the Y phylogenetic tree; the experimental analysis procedures for Y-SNPs should be readily established and widely published. Among the 18 SNPs, 7 were trunk Y-SNPs, the remaining 11 were O-haplogroup downstream markers. The 7 trunk markers, C-M130, D-M174, J-M304, N-M231, O-M175, Q-M242, and R-M207 were used to determine the main branch attribution of a sample on the phylogenetic tree; and 11 downstream O markers, M119, M268, M95, M176, M122, M324, P201, IMS-JST002611, M7, M134 and M117, were used to determine the main subbranch attribution of a sample under haplogroup O. Seven trunk Y-SNPs were analysed using the enzyme digestion method (the seven endonucleases used are listed in Table 1). Two O downstream SNPs (M134 and M117) were analysed using SSP, and nine other O-specific Y-SNPs were analysed with SNaPshot after multiplexed PCR. The primers used, including the multiplex

PCR primers and single-base extension primers for SNaPshot, are listed in Table 1.

Whole Y Sequencing

Fifty-five DNA samples were randomly selected from the 205 samples, with the only requirement that sufficient residual DNA be left after previous enzyme digestion, SSP, or SNaPshot. NGS on the 55 samples were performed on an Illumina HiSeq2000 platform (San Diego, CA, USA). Several bait libraries were created to capture the sequences in a region covering about 11 Mbp on the Y chromosome (Yan et al. 2014). Other steps prior to NGS adopted procedures previously described by Yan and colleagues, including DNA fragmentation, ligation of adaptors, running gel electrophoresis, preparing libraries, designing baits and performing liquid-phase capture (Yan et al. 2014). NGS data were also analysed with standard methods (bwa + SAMtools) (Li et al. 2009; Li and Durbin 2010). Trimmomatic (version 0.36) was adopted to remove low-quality reads and sequencing adapters from the raw reads (Bolger et al. 2014). Duplicate reads were flagged using Picard (version 2.20.8) (MIT 2019). The clean reads were aligned to the human reference genome GRCh37 using the Burrow-Wheeler Aligner (version 0.7.17-r1188) as described by Li et al. 2009, and Li and Durbin 2010. SNPs and small Indels were determined by the Genome Analysis Toolkit version 4 HaplotypeCaller developed by DePristo et al. (2011). Y-SNP and Y-haplogroup names were confirmed with Yleaf 2.0 based on the suggestions from the Y Chromosome Consortium (TSC) (Y Chromosome Consortium 2002) and ISOGG. Especially, we used Yleaf 2.0 (Ralf et al. 2018) to analyse and annotate the haplogroup of each sequenced sample.

Data Analysis and Statistics

For enzyme digestion, SSP, or SNaPshot, the status of each marker was recorded and manually counted, the frequencies were calculated using IBM SPSS Statistics 26, and the haplogroup clades were manually allocated. Chi-square tests were used to compare the differences between the frequencies obtained by the researchers and those from the literature obtained by Lang et al. (2019) using the same low-resolution methods. Principal component analysis (PCA) and multidimensional analysis (MDS) were performed on the data and data from other nine Han populations obtained from published reports (Zhong 2010; Gao 2013) using IBM SPSS Statistics 26. MEGA-X (version 10.2.4) and Interactive Tree Of Life (iTOL) (version 5.7) were employed to construct a neighbour-joining (NJ) tree for the data obtained for ten Han subpopulations (Kumar et al. 2018; Letunic and Bork 2019), which included Hans from Shaanxi, Shandong, and Anhui (around Henan, most regions close to the Yellow River), Heilongjiang (northeast China), Yunnan, Sichuan, Chongqing, and Guizhou (southwest China).

For whole Y sequencing and analysis of the haplogroups reported by Yleaf 2.0, the frequencies of each clade were calculated with IBM SPSS Statistics 26. Bayesian evolution and age estimation (Lu et al. 2020) were analysed by BEAST (v2.0.0) (Bouckaert et al. 2014); for the calibration of age estimation (Karmin et al. 2015), 41,900 years of age was considered for NO-M214 (95% CI = 40,175-43,359 years). Using MegaX, a Bayesian skyline plot (BSP) was performed to further assess passed changes in effective population size (Ne). For this analysis, BEAST v1.8.0 was used to reveal the magnitude and timing of past changes in the population size of male Henan Han individuals; the Y chromosome NGS datasets were analysed with the GTR model; and the clock model was constructed with a mean mutation rate of 3×10^{-8}/site/generation, as used in LAMARC. The analysis was performed over a period of 10 million generations, with parameters recorded every 1000 generations. Tracer (version 1.6) (http://tree.bio.ed.ac.uk/software/tracer/) was employed to check chain convergence as well as to perform the skyline reconstruction.

Additionally, a branching tree was constructed for the haplogroups based on the 55 sequenced samples using Excel, with reference to https://www.yfull.com/.

RESULTS

Y-STR Results

Two hundred and four unique Y-STR haplotypes were detected among the 205 samples.

Table 1: Information on the 18 Y-SNPs identified via traditional low-resolution analysis (enzymatic digestion, SSP, or SNaPshot)

Locus	*RS*	*Mutation*	*Primer sequencesÿ 5→3'*	*Amplicon size*	*Further analysis or sequencing primer*	*Length*
M175		-5 bp	For: ttgagcaagaaaaatagtaccca Rev: ctccattcttaactatctcaggga	444 bp	Go to enzyme digestion with EarI, cutting site (5→3') is: CTCTTC▼, reaction temperature is: 37°C	
M130	35284970	C→T	For: gagtgggagggactgtgaga Rev: ccacagagatggtgtgggta	528bp	Go to enzyme digestion with BslI, cutting site (5→3') is: CCNNN▼NNNGG, reaction temperature is: 55 °C	
M174	2032602	T→C	For: acatctcagatcgttgtttggt Rev: aaaaagccatgcaattacctg	219 bp	Go to enzyme digestion with BsrI, cutting site (5→3') is: ACTGGN▼, reaction temperature is: 65°C	
M231	9341278	G→A	For: cctattatcctggaaaatgtgg Rev: attccgattcctagtcacttgg	331 bp	Go to enzyme digestion with Taq^{á}I, cutting site (5→3') is: T▼CGA, reaction temperature is: 65°C	
M304	13447352	A→C	For: ctagagggtattggggtaggc Rev: ggaaataccttcacctaaata	310 bp	Go to enzyme digestion with Tsp45I, cutting site (5→3') is:▼GTSAT, reaction temperature is: 65°C	
M242	8179021	C→T	For: aactcttgataaaccgtgctg Rev: tccaatctcaattcatgcctc	366 bp	Go to enzyme digestion with Bsp1286I, cutting site (5→3') is: GDGCH▼C, reaction temperature is: 37°C	
M207	2032658	A→G	For: aggaaaaatcagaagtatccctg Rev: caaaattcaccaagaatccttg	423 bp	Go to enzyme digestion with DraI, cutting site (5→3') is: TTT▼AAA, reaction temperature is: 37°C	
M134	200461450	C→del	For: agaatcatcaaacccagaagg Rev: tctttggcttctctttgaacag	231 bp	Go to sequence specific PCR (SSP)	
M117	759551978	-4 bp	For: aagtatgacttatgaagtacgaagaaa Rev: attcagttagattttacaatgagca	429 bp	Go to sequence specific PCR (SSP)	
M119	72613040	A→C	For: acccagttctttcacatagctta Rev: ctcaccctaaggaagtcacga	510 bp	TCGCTTCCAGTCGGATCCATCGTA CGTCTTATGGGT TATTCCAATTC AGCATACAGGC	58 bp
M268	13447443	A→G	For: gttattatccctgtacctgacta Rev: ccggccactatacttcttt	379 bp	ATCGTTCGTCATGCCTAGCCTCATTCCTCT AAAAT	35 bp
M95	2032650	C→T	For: gggagtggaaatcaagatg Rev: agacaatgagggtgggtg	368 bp	ACGTCTAACCTGAGGATAAGGAAAGACTAC CATA TTAGTG	40 bp
M176	11575897	C→T	For: aagccacagtgtccgtttgt Rev: gtatcgacctcgtcggaagg	514 bp	ATCGCTTCCAGTCGGATCCATCGTTTCCCG CTTCG GTACT CTGCAG	46 bp
M122	78149062	T→C	For: gtgatcaacttctttccctcaaca Rev: atgggtttttcttcctttctacca	518 bp	GCATTACGGAGCAATTGAGATACTAATT CA	30 bp
M324	13447361	G→C	For: gggagtcatctggtgtgtaaga Rev: gcagatgggaagtgaaacaacc	654 bp	ATCGCTTCCAGTCGGATCCATCGTATCTGA TTT GATCTACC TGCCCTTTCCT	52 bp
P201	2267801	T→C	For: tccggtgaggtagtgagagtg Rev: ctcctgttcccactaacccc	379 bp	GATCATGACGTCAATGAAGGTAGAAGGGTGA GA GCCAG TTAAAGCCC	47 bp
IMS-JST002611	2075181	C→T	For: atgggtttttcttcctttctacca Rev: ccctagacacttgggtcagc	637 bp	ATCGATCGATCGATCGATCGATCGATCGAT CGAT CG	
M7	3898	C→G	For: ggcatgggaatagacattgagg Rev: tgcctttgatggttgagtgc	592 bp	ATTCCTCTCTGT ATGTCAGATCTAACAA	64 bp

Source: Authors

Population genetic parameters at both haplotype and locus levels for the 205 Han individuals based on the 29 Y-STR data are listed in Table 2 and Table 3. At the haplotype level, the fraction of unique haplotypes was 0.9902, the discrimination capacity (DC) was 0.9951, and the gene diversity (HD) was 1.0000. At the locus level, the average gene diversity (GD) across loci was 0.7070, with the highest GD occurring at 0.8914 in DYS385_b and the lowest GD occurring at 0.4106 in DYS391. No difference was observed between this study's data and the data of 1434 samples cited from the literature (Wang et al. 2016) in terms of unique haplotypes (χ^2 test), fraction of unique haplotypes, discrimination capacity, HD at the haplotype level (ANOVA test) or GD at the single locus level (ANOVA test) (Supplementary Tables S1 and S2) due to all $p>0.05$.

Table 2: Population genetic parameters at the haplotype level for 205 Han people in Henan Province based on Microreader™ 29 Y-STR data

Haplotypes	*Han [Henan]*
Sample size	205
Number of haplotypes	204
Unique haplotypes	203
Fraction of unique haplotypes	0.9902
Discrimination capacity	0.9951
Haplotype-level Gene diversity (HD)	1.0000
(HD) +/- SD	0.0005

Source: Authors

Table 3: Population genetic parameters at the locus level for 205 Hans in Henan Province based on Microreader™ 29 Y-STR data

Locus	*Gene diversity*
DYS456	0.5880
DYS389I	0.6215
DYS390	0.6806
DYS389II	0.7645
DYS458	0.8036
DYS19	0.7210
DYS385_a	0.7630
DYS385_b	0.8914
DYS393	0.5952
DYS391	0.4106
DYS439	0.6643
DYS635	0.7380
DYS392	0.7048
DYS437	0.5079
DYS438	0.5037
GATA_H4	0.6311
DYS448	0.7252
DYS460	0.6693
DYS449	0.8794
DYS576	0.7364
DYS627	0.8401
DYS518	0.8821
DYS570	0.8144
DYS481	0.8098
DYF387S1_1	0.8048
DYF387S1_2	0.8041
DYS533	0.6354
DYS549	0.6209
DYS643	0.6902
Average gene diversity over loci (GD)	0.7070
(GD) +/- SD	0.1190

Source: Authors

Results of Traditional Low-resolution Analysis Techniques

In this study, the haplogroups O, C, D, J, N, Q and R were observed among 205 male Henan Han individuals after enzyme digestion, SSP, or SNaPshot analysis. Among these, the O haplogroup accounted for 75.11 percent. O2 and its downstream sub-branches were observed at the highest frequency (125/205=0.6098) compared to other O subclades. The detailed ratios are presented in Table 4, where C-M130 accounted for 11.71 percent, O-M175 accounted for 75.11 percent, O2a1b-002611 accounted for 17.07 per-

Table 4: Results of traditional low-resolution analysis of the 18 Y-SNPs in the haplogroup of 205 male Henan Han people

Haplogroup	*Number*	*Frequency*
C-M130	24	0.1171
D-M174	1	0.0049
J-M304	2	0.0098
N-M231	9	0.0439
Q-M242	13	0.0634
R-M207	2	0.0098
O-M175	154	0.7511
O1a-M119	16	0.0780
O1b-M268	8	0.0390
O1b1a1a-M95	5	0.0244
O1b2-M176	0	0
O2-M122	1	0.0049
O2a-M324	7	0.0341
O2a1b-002611	35	0.1707
O2a2-P201	17	0.0829
O2a2a1a2-M7	5	0.0244
O2a2b1-M134	24	0.1171
O2a2b1a1-M117	36	0.1756

Source: Authors

cent, O2a2b1-M134 accounted for 11.71 percent, O2a2b1a1-M117 accounted for 17.56 percent, etc. The mutation list for all 205 samples is shown in Supplementary Table S3. No significant difference in Y haplogroup types among male Henan Han individuals was detected after the Chi-square test between present study and that of a 2019 study by Lang et al. (Table 5, χ^2=5.252, p=0.5119). The subclades under O were also further compared, and no significant difference was observed between this study's results and those of a previous study by Lang et al. in 2019 (Table 5, χ^2= 0.0152,p=0.9025).

Table 5: Comparison of Y haplogroup distributions between our 205 haplogroups and those of a previous study (Lang et al. 2019) on male Henan Han people

Haplogroup	*Lang et al. n=199*	*Present study, n=205*
C	31	24[a]
D	4	1 [a]
J	1	2 [a]
N	13	9 [a]
Q	13	13 [a]
R	3	2 [a]
O	134	154 [a]
O1	*26*	*29*[b]
O2	*108*	*125*[b]

Note: a: df=6, χ^2=5.252, p=0.5119; b: df=1, χ^2= 0.0152, p=0.9025
Source: Authors

The Y haplogroup data of ten Han individuals, including data from this study's samples (Henan Hans, or HNH), data from Anhui Hans (AHH) and 8 other Hans cited from Zhong 2010, and data from Yunnan Hans (YNH) cited from Gao's 2013 thesis, are listed in Supplementary Table S4. No significant difference was observed among these populations. The corresponding diagrams are shown in Figures 1, 2 and 3. The PCA, MDS, and NJ tree showed that the northern Han population (SAXH, HNH, HLJH and AHH) was more closely related to the southern Han population (CQH, GZH, YNH, SCH and FJH), and vice versa.

NGS Results

Only the C and O haplogroups were observed among the 55 samples after whole Y sequencing (Supplementary Table S5), which showed a pattern different from that obtained from the 205 samples (Supplementary Table S3). With highly detailed haplogroups obtained from the 55 sequenced samples defined by Yleaf 2.0, the researchers observed that the frequencies of many O subclades (O1a, Olalalala2, Olala2c1, O1bla1b1, O1bla2al, O1bla2b, O1b1a2c, O2a1b1a2a1b, O2a1b1b, O2a2a2a, O2a2b1a2a, O2a2b2b1b, O2b1a, etc.) presented only once and were at a low frequency among modern male Henan Han individuals (each at 1/55=1.82%), whereas four lineages, namely, O2a2b1a (6/55=10.91%), O2a2b1a2a (7/55=12.73%), O2a1a1a1a1 (8/55=14.55%), and O2a1b1a2a1a1a1 (4/55=7.27%), had relatively higher frequencies than the above singletons.

The BSP (Fig. 4) showed that a population increase from 20,000 to 800,000 was observed in Henan Han males at approximately 6,000-13,000 years before present (YBP) (the period was highlighted), which is accordant with the previous report by Lu et al. in 2020.

Comparison of Data from Two Types of Haplogroup Determination

The researchers compared the whole Y sequencing/Y leaf results with traditional results (enzyme digestion/SSP/SNaPshot) for each of the 55 samples, as shown in Table S5. The researchers found consistency in the main clade (C or O) determination between the results of traditional methods and the results of NGS, which indicated that the identification of M130 to determine C as well as M175 to determine O was reliable enough. However, in the subclade determinations, only one sample was the same (O1a) between the two groups (1/55=1.82%). The O subclades of the two samples were absolutely different between the two groups (O1 according to the low-resolution methods became O2 according to Yleaf 2.0, and O2 according to the low-resolution methods became O1 according to Yleaf 2.0; difference rate at 2/55 = 3.64%). The remaining 52 samples all yielded more detailed and clearer subclades according to NGS than according to traditional methods (clearer percentage at 52/55 = 94.54%). With NGS, all clade and subclade determinations were clear and more accurate and extended to the far end of each branch. For example, one sample was identified

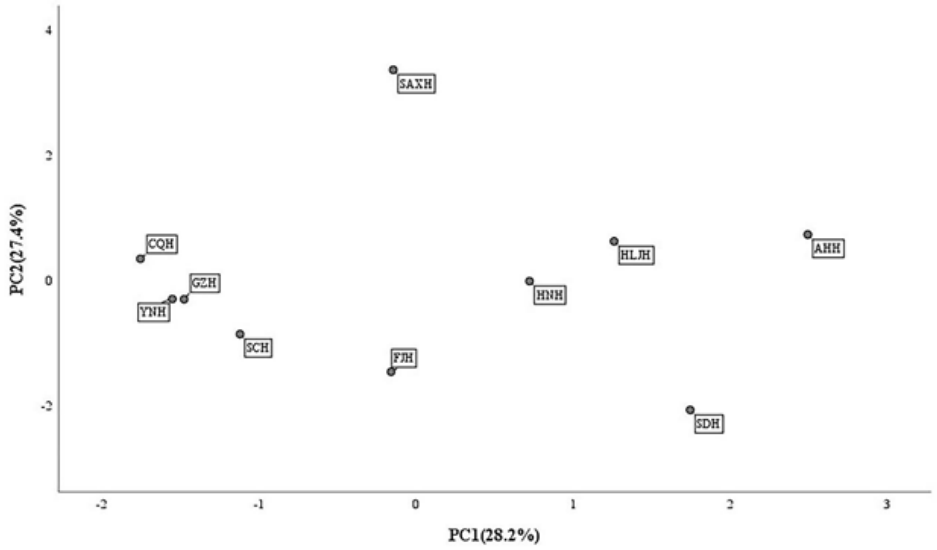

Fig. 1. Principal component analysis of ten Chinese Han populations based on Y-SNP data with SPSS Statistics 26. The data for the other nine Hans were obtained from Zhong (2010)
Population codes: HNH=Henan Han; YNH=Yunnan Han; GZH=Guizhou Han; FJH=Fujian Han; SAXH=Shaanxi Han; SDH=Shandong Han; HLJH=Heilongjiang Han; SCH=Sichuan Han; AHH=Anhui Han; CQH=Chongqing Han
Source: Authors

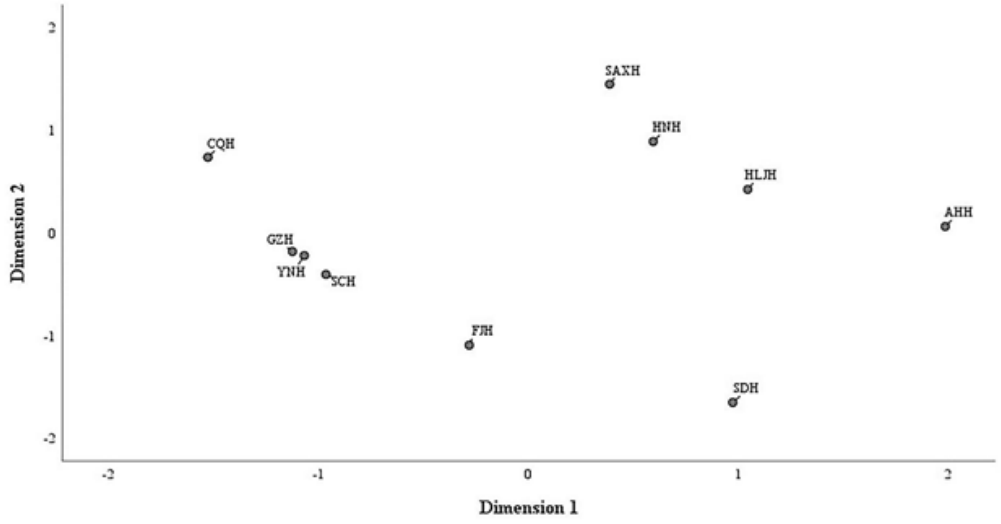

Fig. 2. Multidimensional analysis (MDS) was performed with SPSS 26 with ten Chinese Han populations based on Y-SNP frequencies. The data for the other nine Hans were obtained from Zhong (2010)
The population codes used are shown in Figure 1
Source: Authors

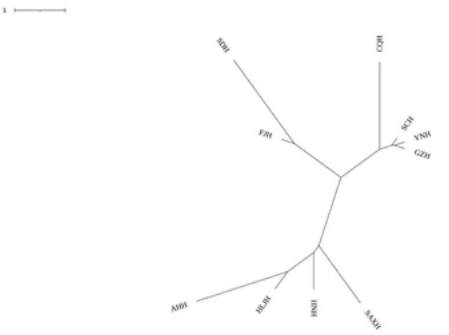

Fig. 3. Neighbor-joining tree of ten Chinese Han populations based on Y-SNP data generated via MEGA-X (version 10.2.4) and interactive tree of life (iTOL) (version 5.7)
The data for the other nine Hans were obtained from Zhong 2010. The population codes used are shown in Figure 1
Source: Authors

to be O1a with low-resolution methods due to the derived O-M119, while it was identified to be O1a1a1a2 by NGS/Yleaf analysis (Table S5). In contrast, the traditional methods yielded vague and less clear subclades than did NGS, and few results obtained by the traditional methods may have been erroneous (approximately 2/55=3.64%).

DISCUSSION

In this study, the researchers first analysed 205 male Henan Han individuals via traditional methods. From the results, the researchers obtained important information about the Y haplogroup pattern of modern Han males in Henan, the starting location of the Han ethnic group. Then, the researchers precisely sequenced 55 Y chromosomes from the 205 samples to evaluate the accuracy and reliability of these traditional methods versus whole Y chromosome sequencing.

By analysing 205 samples via traditional methods, including enzymatic digestion, SSP, or SNaPshot, the researchers obtained a relatively reliable Y haplogroup pattern among Henan Han people supported by enough samples. The researchers found that O was the most prevalent main clade, whereas C, D, J, N, Q, and R were also present. Although He et al. recently reported that O2a2, O2a1, O1b1, O1a1, C2b1, C2a1, O2b1 and Q1a1 were the dominant Y-chromosomal lin-

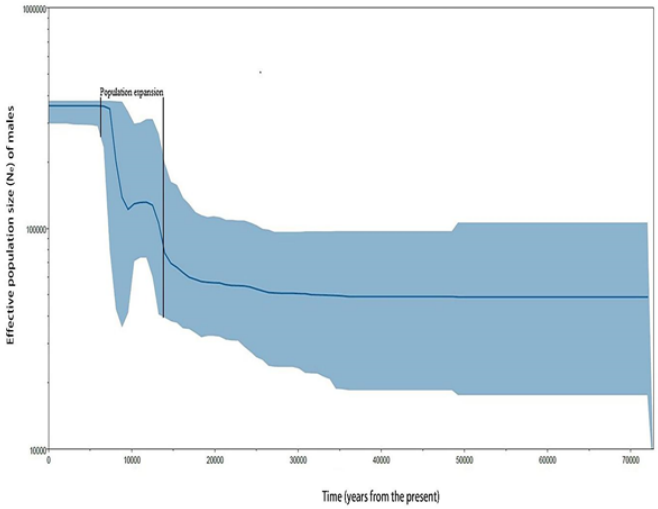

Fig. 4. Bayesian skyline plot (BSP) constructed from the NGS data of 55 male Henan Han individuals
A graph was constructed based on the raw sequencing data of 55 male Henan Han individuals. Changes in the effective population size (Ne) of male Henan Han individuals were detected through 25 years and a mutation rate of 3×10^{-8}/site/generation. The present day is on the left of the X-axis
Source: Authors

eages among Han Chinese individuals in their study of 1033 male Chinese individuals with their newly developed 639-plex Y-SNP panel (He et al. 2023), it is more commonly accepted that D, J, N and R also exist in male Hans (Xue et al. 2008; Wang and Li 2013; Liu et al. 2022; Li et al. 2023; Tao et al. 2023). The researchers' results are also in consistent with those of the 2019 study by Lang et al.

According to this study's PCA, MDS and NJ analyses, Heilongjiang Han (HLJH) was more closely related to HNH than were Shaanxi Hans (SAXH) and Shandong Hans (SDH) to HNH. This finding differs from a number of other studies (Chen et al. 2009; Wang et al. 2021), the conclusions of which, stating that SAXH and SDH are genetically closer to HNH than to HLJH, have been publicly accepted. Here, the researchers assume that the error was a result of the data used (Zhong 2010), which contained an insufficient sample size. For example, Zhong's data included data from only 52 Shaanxi Hans, 52 Shandong Hans and 67 Heilongjiang Hans (Zhong 2010). This study's sequencing of 55 male Henan Han people revealed that only C and O were present, within which D, J, N, Q, and R were not observed. Thus, it was clearly demonstrated that a small sample size can result in a less comprehensive haplogroup distribution and generate biassed results in such studies. Zhong's data were very compatible with this study's data in columns (or Zhong's data also listed C, D, J, N, O, Q and R), as possible comparisons.

The researchers' comparison of the results of low-resolution methods and NGS showed differences between these two kinds of techniques (traditional versus NGS), indicating that NGS/Yleaf can considerably enhance sub-branch determinations, permitting the identification of the far end of the O haplogroup, for example, from O1a to O1a1a1a1a2. This factor leads to a question applicable to previous reports based on traditional methods (including enzyme digestion, SSP, or SNaPshot), which is whether those previous haplogroup results can be considered accurate and reliable enough in subclade determinations (as mentioned in Butler 2003). Therefore, many previous low-resolution studies on subclade determinations should be considered less reliable and outdated. Fortunately, traditional methods for determining the main clades are highly reliable, for example, M130 is used to determine C, and M122, M119, and M324 are used to determine O, and these results were confirmed by NGS in this study. The researchers found that the main clade determinations on the same sample either by traditional methods or by NGS were completely consistent (to be the C haplogroup or to be the O haplogroup), although these two kinds of methods could generate different subclade allocations for the same sample.

The present research, based on Y sequencing data, revealed high diversity of paternal lineages among male Henan Han individuals. The sequencing of 55 Y chromosomes from male Henan Hans also revealed that four lineages (O2a2b1a1a, O2a2b1a2a, O2a1b1a1a1 and O2a1b1a2a1a1) contributed more significantly to the gene pool of modern Hans than to that of other lineages, and male Henan Han individuals also experienced a remarkable population expansion in the Neolithic Age (that is, approximately 9000 YBP). These findings were in accordance with the results of other studies. In 2014, Yan et al. reported that 40 percent of China's present Han population descended from three branches (Oa-F5, Ob-F46, Oc-F325). Coincidently, Lu et al. also reported the rapid expansion of several branches (Oa-F5, Ob-F46, Oc-F325, and O2a1b1b-F1273) during the Neolithic age (9000 YBP) in this area, which may have contributed to the large proportion of Henan Han population (Lu et al. 2020). He et al. recently reported that O2a2, O2a1, O1b1, O1a1, C2b1, C2a1, O2b1 and Q1a1 were the major Y-chromosomal lineages among Han Chinese individuals (He et al. 2023). Thus, the present study provides valuable insights regarding the origin and admixture of modern Hans and their most recent ancestor, the 'Huaxia' population.

Regarding Han's origin, Su et al. in 1999 proposed a scenario indicating southern origin supported by genetic evidence from the M214/M175 lineage, suggesting that *Homo sapiens* entered East Asia from the south (Su et al. 1999), while other authors have proposed a northern route for human migration to Central and Northeast Asia, suggesting that *Homo sapiens* entered East Asia from the north (Jin and Su 2000; Li et al. 2019; Li et al. 2023). Interestingly, the present-day Henan region may be one of the key sites for the confluence of southern and northern ancestors (Li et al. 2019). However, only a small

number of relevant research papers have been published on Henan Han people in addition to the 2019 study by Lang et al. and this could be because they included relatively fewer Henanese subjects or because they did not have enough genetic markers. Additionally, several large-scale southward population migrations from Henan and the surrounding areas to southern China have been recorded since the 180s A.D. (Wang et al. 2022), which also promoted the dissemination and intermixing of Hans from Henan to other regions, complicating the formation process of modern Hans (Yu and Li 2021). Undoubtedly, the early inhabitants of the Henan area contributed significantly to the formation of the modern Chinese Han population. Therefore, the study of 205 samples and subsequent sequencing of the Y chromosomes of 55 male Henan Han individuals contribute to revealing the genetic components of Henan Han males as well as clarifying the origin and admixture of the Han ethnic group.

CONCLUSION

This study provides basic evidence that traditional techniques, such as enzyme digestion, sequence-specific PCR, and SNaPshot NGS, can be used to determine the main branch of Y haplogroups but are less accurate and less reliable in sub-branch determinations than NGS/Yleaf is (with a difference rate of 3.64% and a less clear rate of 94.54%). This study demonstrated that NGS/Yleaf yields clearer subclade determinations than traditional techniques and can allocate a sample to the far terminus of a haplogroup. This study also revealed that Y haplogroups O, C, D, J, N, O, Q and R existed among the male Hans in Henan Province, with O accounting for 75.11 percent and C accounting for 11.71 percent. The question of how those Y haplogroup results generated by traditional techniques can be further utilised remains to be addressed.

RECOMMENDATIONS

Future studies with large-scale samples analysed by NGS/Yleaf should be performed to determine additional discrepancies between traditional methods and NGS/Yleaf in Y haplogroup determinations, as these studies will help to elucidate the Y haplogroup pattern of male Henan Han individuals.

ABBREVIATIONS

B.C.E.: Before Common Revolution
BSP: Bayesian Skyline Plot
C.E.: Common Era
DC: discrimination capacity
GD: gene diversity
MDS: multidimensional analysis
NGS: next-generation sequencing
NRY: non-recombining Y (chromosome region)
PCA: Principal Component Analysis
SNPs: single nucleotide polymorphisms
SSP: sequence-specific PCR
STR: short tandem repeat
YBP: years before present

SUPPLEMENTARY INFORMATION

Comparisons based on 27 Y-STRs between the 205 samples and 1434 previously reported male Henan Han samples (Wang et al. 2016) at the haplotype level are listed in Supplementary Table S1. Comparisons of the genetic diversity at the single-locus level between the 205 Y-STRs and previously reported 1434 male Henan Han samples (Wang et al. 2016) are listed in Supplementary Table S2. The Y-SNP mutations identified in the 205 samples via traditional analysis methods are presented in Supplementary Table S3. Y haplogroup data of male Hans from ten provinces in China are listed in Supplementary Table S4. The results of next-generation sequencing and low-resolution methods on 55 randomly selected male Henan Han individuals are presented in Supplementary Table S5.

DATA ACCESS

For this study, the raw sequence data has been stored to the GSA (Wang et al. 2017) at the Beijing Institute of Genomics Data Centre (BIG Data Center Members 2017) of Chinese Academy of Sciences. The accession number is HRA000062, which can be freely visited at https://ngdc.cncb.ac.cn/gsa-human/browse/HRA000062. The raw sequence data were also submitted to NCBI, which can be freely accessed and downloaded at https://www.ncbi.nlm.nih.gov/bioproject/PRJ-

NA867000. The script for data analysis has also been uploaded to GitHub at https://github.com/wanghao958876/data-analysis.

ACKNOWLEDGEMENTS

The researchers thank all the donors for their participation and cooperation in this study. The researchers thank Dr. Qi LU at the Key Laboratory of Forensic Genetics, Institute of Forensic Science, Ministry of Public Security, China, and Dr. Hui-Zhen Chen at the Department of Anthropology and Ethnology, Institute of Anthropology, Xiamen University, for providing kind help with this study. Support for the data analysis was provided by the Supercomputing Centre of Zhengzhou University.

FUNDING

This study was supported by the National Natural Science Foundation of China (U1804194, 31900406 and 82101963), the Scientific and Technology Committee of Shanghai Municipality (18490750300) and the Fundamental Research Funds for the Central Universities (20720191047).

AUTHORS' CONTRIBUTIONS

Hao Wang, Lu Yu and Ning Wang carried out the studies and participated in sample collection, DNA extraction and data analysis, and they were all first authors. Lulu Wang and Zhongqian Guo completed the remaining experimental operations. Lanhai Wei, Hong'en Xu and Zhaoshu Zeng all designed the experiments, wrote the article and discussed the manuscript. All the authors edited and approved the final manuscript.

REFERENCES

BIG Data Center Members 2017. The BIG Data Center: from deposition to integration to translation. *Nucleic Acids Res,* 45(D1): D18–D24.

Bolger AM, Lohse M, Usadel B 2014. Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics*, 30(15): 2114-2120.

Bouckaert R, Heled J, Kühnert D et al. 2014. BEAST 2: A software platform for Bayesian evolutionary analysis. *PLoS Comput Biol,* 10(4): e1003537.

Butler JM 2003. Recent developments in Y-Short Tandem repeat and Y-Single Nucleotide polymorphism analysis. *Forensic Sci Rev*, 15(2): 91-111.

Chen J, Zheng H, Bei JX et al. 2009. Genetic structure of the Han Chinese population revealed by genome-wide SNP variation. *Am J Hum Genet,* 85(6): 775-785.

Chen P, Wu J, Luo L et al. 2019. Population genetic analysis of modern and ancient DNA variations yields new insights into the formation, genetic structure, and phylogenetic relationship of Northern Han Chinese. *Front Genet* 10: 1045.

DePristo MA, Banks E, Poplin R et al. 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet,* 43(5): 491-498.

Excoffier L, Lischer HE 2010. Arlequin suite ver 3.5: A new series of programs to perform population genetics analyses under Linux and Windows. *Mol Ecol Resour,* 10(3): 564-567.

Gao J 2013. Genetic Polymorphism And Forensic Implications Of Y-SNP In Yunnan Han Population. Master Thesis. Yunnan: Kunming Medical University. From <https://kns.cnki.net/kcms2/article/abstract?v=3uoqIhG8C475KOm_z rgu4lQARvep 2SAk8URRK9V8kZLG_vkiPpTeIQJhOZ yUDXmqpZ c8YdI1DZ lvvyxvPtuPyvd ZoGZ9DP 49& uniplatform= NZKPT> (Retrieved on 6 February 2024).

He G, Wang M, Miao L et al. 2023. Multiple founding paternal lineages inferred from the newly-developed 639-plex Y-SNP panel suggested the complex admixture and migration history of Chinese people. *Hum Genomics* 17(1): 29.

ISOGG 2020. Y-DNA Haplogroup Tree 2019, Version: 15.73. From <http://www.isogg.org/tree/> (Retrieved on 6 February 2024).

Jin L, Su B 2000. Natives or immigrants: Modern human origin in East Asia. *Nat Rev Genet,* 1(2): 126-133.

Karafet TM, Mendez FL, Meilerman MB et al 2008. New binary polymorphisms reshape and increase resolution of the human Y chromosomal haplogroup tree. *Genome Res*, 18(5): 830-838.

Karmin M, Saag L, Vicente M et al. 2015. A recent bottleneck of Y chromosome diversity coincides with a global change in culture. *Genome Res,* 25(4): 459-466.

Kumar S, Stecher G, Li M et al. 2018. MEGA X: Molecular evolutionary genetics analysis across computing platforms. *Mol Biol Evol,* 35(6): 1547-1549.

Lang M, Liu H, Song F et al. 2019. Forensic characteristics and genetic analysis of both 27 Y-STRs and 143 Y-SNPs in Eastern Han Chinese population. *Forensic Sci Int Genet,* 42: e13-e20.

Letunic I, Bork P 2019. Interactive Tree Of Life (iTOL) v4: Recent updates and new developments. *Nucleic Acids Res*, 47(W1): W256-W259.

Li F, Vanwezer N, Boivin N et al. 2019. Heading north: Late Pleistocene environments and human dispersals in central and eastern Asia. *PLoS One,* 14(5): e0216433.

Li H, Durbin R 2010. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics,* 26(5): 589-595.

Li H, Handsaker B, Wysoker A et al. 2009. The sequence alignment/map format and SAMtools. *Bioinformatics,* 25(16): 2078-2079.

Li J, Song F, Lang M et al. 2023. Comprehensive insights into the genetic background of Chinese populations using Y chromosome markers. *R Soc Open Sci,* 10(9): 230814.

Liu J, Jiang L, Zhao M et al. 2022. Development and validation of a custom panel including 256 Y-SNPs for Chinese Y-chromosomal haplogroups dissection. *Forensic Sci Int Genet,* 61: 102786

Liu L 2005. *The Chinese Neolithic: Trajectories to Early States.* Cambridge, England: Cambridge University Press.

Lu Q, Cheng HZ, Li L et al. 2020. Paternal heritage of the Han Chinese in Henan province (Central China): High diversity and evidence of in situ Neolithic expansions. *Ann Hum Biol,* 47(3): 294-299.

MIT 2019. Picard (version 2.20.8). From <http://broadinstitute.github.io/picard/> (Retrieved on 6 February 2024).

Poznik GD, Xue Y, Mendez FL et al. 2016. Punctuated bursts in human male demography inferred from 1,244 worldwide Y-chromosome sequences. *Nat Genet,* 48(6): 593-599.

Ralf A, Montiel González D, Zhong K et al. 2018. Yleaf: Software for human Y-chromosomal haplogroup inference from next-generation sequencing data. *Mol Biol Evol,* 35(5): 1291-1294.

Shi H, Dong YL, Wen B et al. 2005. Y-chromosome evidence of southern origin of the East Asian-specific haplogroup O3-M122. *Am J Hum Genet,* 77(3): 408-419.

Su B, Xiao J, Underhill P et al. 1999. Y-Chromosome evidence for a northward migration of modern humans into Eastern Asia during the last Ice Age. *Am J Hum Genet,* 65(6): 1718-1724.

Tao R, Li M, Chai S, et al. 2023. Developmental validation of a 381 Y-chromosome SNP panel for haplogroup analysis in the Chinese populations. *Forensic Sci Int Genet,* 62: 102803.

Twitchett D, Loewe M 1986. *The Cambridge History of China, Vol. 1: The Ch'in and Han Empires, 221 BC-AD 220.* Cambridge, England: Cambridge University Press.

Wang CC, Li H 2013. Inferring human history in East Asia from Y chromosomes. *Investig Genet*, 4(1): 11.

Wang L, Chen F, Kang B et al. 2016. Genetic population data of Yfiler Plus kit from 1434 unrelated Hans in Henan Province (Central China). *Forensic Sci Int Genet,* 22: e25-e27.

Wang M, Yuan D, Zou X et al. 2021. Fine-scale genetic structure and natural selection signatures of Southwestern Hans inferred from patterns of genome-wide allele, haplotype, and haplogroup lineages. *Front Genet,* 12: 727821.

Wang Y, Song F, Zhu J et al. 2017. GSA: Genome sequence archive. *Genomics Proteomics Bioinformatics,* 15(1): 14-18.

Wang Y, Zou X, Wang M et al. 2022. The genomic history of southwestern Chinese populations demonstrated massive population migration and admixture among proto-Hmong-Mien speakers and incoming migrants. *Mol Genet Genomics*, 297(1): 241-262.

Xue F, Wang Y, Xu S et al. 2008. A spatial analysis of genetic structure of human populations in China reveals distinct difference between maternal and paternal lineages. *Eur J Hum Genet,* 16(6): 705-717.

Yangshao Culture 2020. New World Encyclopedia. From <https://www.newworldencyclopedia.org/p/index.php?title=Yangshao_culture&oldid=1032252> (Retrieved on 25 February 2024).

Yan S, Wang CC, Zheng HX et al. 2014. Y chromosomes of 40% Chinese descend from three Neolithic super-grandfathers. *PLoS One,* 9(8): e105691.

Y Chromosome Consortium 2002. A nomenclature system for the tree of human Y-chromosomal binary haplogroups. *Genome Res*, 12(2): 339-348.

Yu L 2018. Y-SNP Genetic Polymorphisms of 209 Men in Henan. Master Thesis. Henan: Zhengzhou University. From <https://kns.cnki.net/kcms/detail/detail.aspx? dbcode=CMFD&dbname=CM FD201802& filename= 10180 93184.nh&v=nG7 WTv9EDWJ3Jn OJVVM 8ikIclW4Bu1 MWnNdcA1aem06xgjzrq9 rlEmRAp3 msksJi> (Retrieved on 6 February 2024).

Yu X, Li H 2021. Origin of ethnic groups, linguistic families, and civilizations in China viewed from the Y chromosome. *Mol Genet Genomics* 296(4): 783-797.

Zhong H 2010. Y Chromosome Population Structure and Prehistoric Migrations of East Asians. PhD Thesis. Beijing: Chinese Academy of Sciences. From <http://d. wan fangdata.com.cn/thesis/Y1856586.>_(Retrieved on 6 February 2024).

Zhong H, Shi H, Qi XB et al. 2011. Extended Y chromosome investigation suggests postglacial migrations of modern humans into East Asia via the northern route. *Mol Biol Evol,* 28(1): 717-727.

Paper received for publication in September, 2023
Paper accepted for publication in March, 2024