Insights From Y-STRs: Forensic Characteristics, Genetic Affinities, and Linguistic Classifications of Guangdong Hakka and She Groups

Chunfang Luo1,2†, Lihong Duan3†, Yanning Li1,4†, Qiqian Xie1†, Lingxiang Wang5, Kai Ru5, Shahid Nazir6, Muhammad Jawad6, Yifeng Zhao7, Fenfen Wang8, Zhengming Du8, Dehua Peng8, Shao-Qing Wen5*‡, Pingming Qiu1*‡ and Haoliang Fan1,5,9*

1 School of Forensic Medicine, Southern Medical University, Guangzhou, China, 2 Heyuan Municipal Public Security Bureau, Heyuan, China, 3 Beijing Municipal Public Security Bureau, Beijing, China, 4 School of Basic Medicine, Gannan Medical University, Ganzhou, China, 5 Institute of Archaeological Science, Fudan University, Shanghai, China, 6 Department of Forensic Sciences, University of Health Sciences, Lahore, Pakistan, 7 Nanjing Zhenghong Judicial Identification Institute, Nanjing, China, 8 First Clinical Medical College, Hainan Medical University, Haikou, China, 9 School of Basic Medicine and Life Science, Hainan Medical University, Haikou, China

Guangdong province is situated in the south of China with a population size of 113.46 million. Hakka is officially recognized as a branch of Han Chinese, and She is the official minority group in mainland China. There are approximately 25 million Hakka people who mainly live in the East and North regions of China, while there are only 0.7 million She people. The genetic characterization and forensic parameters of these two groups are poorly defined (She) or still need to be explored (Hakka). In this study, we have genotyped 475 unrelated Guangdong males (260 Hakka and 215 She) with Promega PowerPlex® Y23 System. A total of 176 and 155 different alleles were observed across all 23 Y-STRs for Guangdong Hakka (with a range of allele frequencies from 0.0038 to 0.7423) and Guangdong She (0.0047–0.8605), respectively. The gene diversity ranged from 0.4877 to 0.9671 (Guangdong Hakka) and 0.3277–0.9526 (Guangdong She), while the haplotype diversities were 0.9994 and 0.9939 for Guangdong Hakka and Guangdong She, with discrimination capacity values of 0.8885 and 0.5674, respectively. The gene diversity ranged from 0.4877 to 0.9671 (Guangdong Hakka) and 0.3277–0.9526 (Guangdong She), while the haplotype diversities were 0.9994 and 0.9939 for Guangdong Hakka and Guangdong She, with discrimination capacity values of 0.8885 and 0.5674, respectively. With reference to geographical and linguistic scales, the phylogenetic analyses showed us that Guangdong Hakka has a close relationship with Southern Han, and the genetic pool of Guangdong Hakka was influenced by surrounding Han populations. The predominant haplogroups of the Guangdong She group were O2-M122 and O2a2a1a2-M7, while Guangdong She clustered with other Tibeto-Burman language-speaking populations (Guizhou Tujia and Hunan Tujia), which shows us that the Guangdong She group is one of the branches of Tibeto-Burman populations and the Huonie dialect of She languages may be a branch of Tibeto-Burman language families.

Keywords: Guangdong Hakka, Guangdong She, Y-STR, forensic characteristics, phylogenetic analyses
INTRODUCTION

Hakka is one of the far-reaching ethnic groups that have a worldwide distribution and is officially recognized as a branch of Han Chinese in China. Hakka is a unique group that is not named after the region (Zhong, 2019). There are about 80 million Hakka people, and ~50 million are situated in southern parts of China (mainly including Guangdong, Jiangxi, Fujian, Guangxi, Sichuan, Hainan, Hunan, Zhejiang, Taiwan, Hongkong, and Macao). Guangdong province is an important region for Lingnan culture, which has its unique styles of language, history, and culture and which lies in the southernmost part of mainland China. The Hakka population is mainly settled in the eastern and northern regions of Guangdong, comprising Meizhou, Heyuan, Huizhou, Shaoqian, and Qingyuan (Figure 1). The origin of Hakka has not been clearly defined yet. At present, there are two views about the origin of the Hakka population, either they belong to Northern Han (Luo, 1989) or they belong to Southern Han (Fang, 2007). According to a previous study (Li et al., 2003), majority of the Fujian Hakka gene pool (80.2%) came from Northern Han based on 14 Y-SNPs. On the other hand, the frequency of a 9-bp deletion in mitochondrial region V is 21.74% in Meizhou Hakka, which indicated that Meizhou Hakka had close relationships with Fujian Hakka (~0.197) and other populations from South China (Cai et al., 2015). From the perspective of physical anthropology, Zheng et al. found that the physical characteristics of the Hakka population are a mixture of South-Asian and North-Asian ethnic populations, which is determined by 86 anthropologic characteristics in 650 male and 704 female Chinese Hakka adults living in Guangdong and Jiangxi (Zheng et al., 2013). Guangdong province is home to the Hakka population, but there is no comprehensive study available on the Y-chromosomal prospect of Guangdong Hakka.

She is one of the largest ethnic minorities in Jiangxi, Fujian, and Zhejiang provinces. Their presence is also reported in the provinces of Anhui and Guangdong. She people are shifting cultivation group in South China and had migrated from their primitive habitations in Fenghuang Mountain in Guangdong Chaozhou to Fujian, Zhejiang, Jiangxi, Anhui, and other provinces for more than a thousand years (Shi, 2013; Liu et al., 2017). The total population of the She group is 709,592 according to the 2010 census. Their presence in South Zhejiang is 25% and 53% in East Fujian. Their presence in Guangdong province is only 3.8% (27,000), although Guangdong Chaozhou Fenghuang Mountain is believed to be the definite headstream of She group’s culture and civilization (Liu et al., 2017). She people speak the She language with different dialects such as Shanke, Dongjia, and Huonie. She language is a branch of Sino-Tibetan language families and is the primary language for the She group (Lei, 2005). With reference to linguistics, the She language (mainly refers to Shanke and Dongjia dialects) belongs to the Hmong-Mien language branch (Lu and Wang, 2019). The Guangdong She population uses only the Huonie dialect. Interestingly, the divergences in languages may represent the differences in the genetic affinities of distinct She groups. Fujian She is more closely related to the Dongbei (Northern Han) and Hunan (Southern Han) populations (Liu et al., 2017). The genetic heterogeneities for the She population in the different regions indicate that the origins of Guangdong She and Fujian She may not be the same; Guangdong She may also not be a Hmong-Mien language-speaking group.

Y-chromosomal variant analysis for determining the patterns of present and past flow of genes between populations is very helpful (Oppenheimer, 2012). Y-chromosome short tandem repeats (Y-STRs) play an important role in forensic molecular biology (Prinz et al., 1997; Adnan et al., 2016). The use of Y-STRs also allows the simultaneous analysis of closely related and distantly related populations (Ballantyne and Kayser, 2012). Consequently, to address three main issues—(1) the forensic efficiencies of Promega PowerPlex® Y23 System (Promega Corporation, Fitchburg, WI, United States) in Guangdong Hakka and Guangdong She groups, (2) the population structures of the two groups, and (3) the language classification of Guangdong She (Huonie Dialect)—we used the Promega 5-dye multiplex system to genotype 23 Y-STRs in 260 Guangdong Hakka and 215 Guangdong She males and evaluated the system effectiveness of forensic applications in two groups. Furthermore, we conducted population genetics employing diversified analyses among globally dispersed human populations and regional closely related populations to make the issues clearer.

MATERIALS AND METHODS

Sample Preparation

In this study, a total of 475 unrelated male individuals (260 Hakka and 215 She) were recruited from Guangdong province of China (Figure 1C). Blood samples of all volunteers were collected using the FTA cards (WhatmanTM, GE Healthcare, Chicago, IL, United States), with their written informed consent. This study and the procedures were approved by the Institutional Review Boards of Hainan Medical University and the Medical Ethics Committee of Hainan Medical University (no. HYLL-2020-012). All the experimental procedures were performed following the standards of the Declaration of Helsinki.

Y-STR Amplification and Genotyping

One punched bloodstain paper (the size was 1.2 mm × 1.2 mm) was used as a PCR template directly for each sample without DNA extraction procedures. Y-STRs were co-amplified using Promega PowerPlex® Y23 System (Promega Corporation, Fitchburg, WI, United States), which is a five-dye multiplex kit that analyzes the 17 Yfiler Y-STRs (DYS19, DYS385a/b, DYS389I, DYS389II, DYS390, DYS391, DYS392, DYS393, DYS437, DYS438, DYS439, DYS448, DYS456, DYS458, DYS635, and Y_GATA_H4) together with six discriminating Y-STRs (DYS439, DYS448, DYS456, DYS635, and Y_GATA_H4) using the FTA cards (Whatman, Waltham, MA, United States). The amplified products were separated by capillary electrophoresis on a 3500XL Genetic Analyzer (Thermo Fisher Scientific, Waltham, MA, United States), following the manufacturer’s instructions. The amplified products were separated by capillary electrophoresis on a 3500XL Genetic Analyzer (Thermo Fisher Scientific, Waltham, MA, United States), and data were analyzed using GeneMapper® ID-X Software v1.6 (Thermo Fisher Scientific, Waltham, MA, United States).
Statistical Analysis

Haplotype and allele frequencies were calculated using Arlequin v3.5 (Excoffier and Lischer, 2010). Gene diversity (GD) was calculated as per the following formula:

\[ HD = \frac{n}{n-1} \left(1 - \sum_{i=1}^{n} P_i^2 \right) \]

where, \( n \) is the number of alleles at each Y-STR locus, and \( P_i \) is the frequency of the \( i \)-th allele. Haplotype diversity (HD) was estimated similarly as GD, while \( n \) and \( P_i \) were the total number of haplotypes and the frequency of the \( i \)-th haplotype, respectively. Discrimination capacity (DC) was determined as the ratio between the number of different haplotypes and the sample size. Match probability (MP) was defined as \( MP = \sum P_i^2 \), where \( P_i \) was the frequency of the \( i \)-th haplotype.

Genetic relationships between these populations and other reference populations were calculated employing \( R_{ST} \) (Excoffier et al., 1992; Excoffier and Smouse, 1994). Population pairwise genetic distances (\( R_{ST} \)) and corresponding \( p \) values between different populations were estimated by analysis of molecular variance (AMOVA) by the online tool available at YHRD\(^1\). The principal component analysis (PCA) and multidimensional scaling plot (MDS) were conducted based on allele frequencies using the R programming language. Additionally, phylogenetic relationships among different populations were drawn in Molecular Evolutionary Genetics Analysis 7.0 software (Kumar et al., 2016) by neighbor-joining phylogenetic tree (Saitou and Nei, 1987) based upon genetic distance matrices (\( R_{ST} \) matrices) and visualized by the Interactive Tree of Life v5 (Letunic and Bork, 2019).

Quality Control

The Y-STR typing experiments were performed strictly following the recommendations of the Chinese National Standards, Scientific Working Group on DNA Analysis Methods (SWGDAM, 2010) and the recommendations of the DNA Commission of the International Society of Forensic Genetics (Gusmao et al., 2006; Carracedo et al., 2013; Roewer et al., 2020). Control DNA 9948 and sdH2O were employed as positive and negative controls in each batch of PCR amplification and electrophoresis, respectively. In addition, our laboratory has passed the proficiency testing for Y-STR typing which is organized by YHRD and has been accredited under ISO/IEC 17025:2005 and China National Accreditation Service for Conformity Assessment.

The haplotype data of 475 unrelated male individuals from Guangdong Hakka and Guangdong She populations in the present study have been submitted to the YHRD database and received the accession number YA004707 (Guangdong Hakka, \( n = 260 \)) and YA004708 (Guangdong She, \( n = 215 \)). The Y-STR profiles, with variants (null alleles, off-ladder alleles, or copy number variants) for all samples, were re-amplified and re-genotyped by AmpFLSTR® Yfiler® PCR Amplification Kit (Thermo Fisher Scientific, Waltham, MA, United States) and Geno-ID Y41 Human Typing (Guangzhou Koalson Intelligent BioRobotics, Guangzhou, Guangdong, China) for confirmation.

\(^1\)https://yhrd.org/
RESULTS AND DISCUSSION

Allele Frequencies, Allele Numbers, and GD Values

The allele frequencies, allele numbers, and GD values of 23 Y-STR loci in the Guangdong Hakka and She groups are presented in Supplementary Tables 1, 2, respectively.

In Guangdong Hakka, 176 distinct alleles were observed at all loci, and the corresponding allelic frequencies ranged from 0.0038 to 0.7423 (DYS438). The number of different alleles varied from three (DYS437) to 40 (DYS385a/b). The GD values of six Y-STRs were greater than 0.9, and the highest (0.9671) and lowest (0.4877) estimates of GD corresponded to loci DYS385a/b and DYS438, respectively. We identified three off-ladders at DYS576 (19.1) and DYS458 (16.1) and the microvariant 19.1 at DYS576 which happened twice. All loci having intermediate values found in the Guangdong Hakka were re-amplified and re-genotyped for confirmation and were commonly observed in YHRD.

In the Guangdong She group, a total of 155 different alleles were obtained, and the number of diverse alleles ranged from four (at seven Y-STRs) to 36 (at DYS385a/b). The allele frequencies varied from 0.0047 to 0.8605 (DYS438). On one hand, the set of 23 Y-STR loci had a high level of genetic polymorphisms in the She group, DYS458 (0.9526), DYS385a/b (0.9180), DYS481 (0.9146), and DYS448 (0.9060), while the DYS438 (0.3277) and DYS391 (0.4875) showed extreme values of the GD distribution (with GD values < 0.5). In this study, no null allele, off-ladder, and copy number variant were found in the Guangdong She group.

Haplotypes and Forensic Parameters

The haplotypes and haplotype frequencies of 23 Y-STRs in the two groups are shown in Supplementary Tables 3, 4, respectively. A total of 231 different haplotypes were found in 260 Guangdong Hakka male individuals, of which 208 (90.04%) were unique. It was observed that 18 haplotypes (H006–H023) occurred twice, four haplotypes (H002–H005) occurred thrice, while only one haplotype (H001) was observed as being shared by four individuals (Supplementary Table 3). Genotyping with the 23 Y-STRs determined 122 distinct haplotypes in the Guangdong She group, and the fraction of unique haplotypes was only 53.28%. Additionally, 43 haplotypes (S015–S057) were detected twice, six haplotypes (S009–S014) and five haplotypes (S004–S008) were found three and four times, respectively, and the haplotypes S003, S002, S001 appeared five, six, and 15 times, respectively (Supplementary Table 4).

As shown in Table 1, the overall HD values of the Guangdong Hakka and She groups were 0.9994 and 0.9939, respectively. Moreover, it can be seen from Table 1 that the DC value of the Guangdong She group (0.5674) was much lower than those of the Guangdong Hakka group (0.8885) and Guangdong Han group (0.9706), which demonstrated that the discrimination power of Promega PowerPlex® Y23 System was not seemingly appropriate for the Guangdong She group. With the Guangdong She group being relatively isolated, there was limited gene flow, which led to the decrease in the discrimination capacity of the Y-STR system. In this regard, there is a dire need to get more knowledge about the mutability of currently known Y-STRs and to incorporate the rapidly mutating Y-STRs into forensic systems to enhance the capacity of discrimination, especially for isolated groups.

PCA

Dimensionality reduction analyses including PCA, MDS, linear discriminant analysis, Laplacian Eigenmaps, and locally linear embedding can accelerate the speed of algorithm execution, improve the performance of the analysis model, and reduce the complexity of data at the same time. To illustrate the genetic landscapes of globally dispersed human populations, especially for Guangdong Hakka and She groups, the dimensionality reduction analyses (PCA and MDS) were conducted based on the frequencies of 23 Y-STRs.

The PCA provides a method of visualizing the essential patterns of genetic relationships and allows us to identify and plot the major patterns within a multivariate dataset to indicate that the populations with closer geographical distances have more intimate relationships. We collected the 23 Y-STR frequency profiles of 48,162 individuals from 159 worldwide populations having nine language families within five continents (Supplementary Table 5) to conduct PCA with Guangdong Hakka and She groups (Yaju et al., 2014; Westen et al., 2015; Bai et al., 2016b; García et al., 2016; Jung et al., 2016; Pickrahn et al., 2016; Wang L. et al., 2016; Wang Y. et al., 2016; Guo et al., 2017; Iacovacci et al., 2017; Jun et al., 2017; Lacerenza et al., 2017; Qiang et al., 2017; Spolnicka et al., 2017; Wang M. et al., 2017; Zgonjanin et al., 2017; Zongjian et al., 2017; Zhang et al., 2017; Cao et al., 2018; Fan et al., 2018a,b,c; Khubrani et al., 2018; Liu Y. et al., 2018; Lacerenza et al., 2017; Liu et al., 2018; Wang et al., 2018, 2020; Zhou et al., 2018; D’Atanasio et al., 2019; Fan et al., 2019; He et al., 2019; Henry et al., 2019; Ip et al., 2019; Jankova et al., 2019; Kai et al., 2019; Lang et al., 2019; Liu et al., 2019, 2020; Luo et al., 2019; Santos Stange et al., 2019; Tao et al., 2019a,b; Wang C. Z. et al., 2019; Wang X. et al., 2019; Wang Y. et al., 2019; Wu et al., 2019; Xie et al., 2019; Yaju et al., 2019, 2020;
FIGURE 2 | Principal component analysis (PCA) based on the frequencies of 23 Y-STRs among Guangdong Hakka, Guangdong She, and 159 worldwide human populations. (A) PCA from geographical scale. (B) PCA from linguistic scale.

FIGURE 3 | Principal component analysis (PCA) based on the 23 Y-STRs frequency profiles among Guangdong Hakka, Guangdong She, and 83 Chinese populations. (A) PCA from geographical scale. (B) PCA from linguistic scale.

Zhabagin et al., 2019; Zhang et al., 2019; Al-Snan et al., 2020; Dezhi et al., 2020; Fan et al., 2020; Feng et al., 2020; Haiying et al., 2020; Hanguang et al., 2020; Jannuzzi et al., 2020; Kang et al., 2020; Shonhai et al., 2020; Song et al., 2020; Tang et al., 2020; Yin et al., 2020; Zeyad et al., 2020). As shown in Figure 2, the first and second components (PC1 and PC2) accounted for 6.85 and 5.69% of the total variances observed within these populations, respectively. From the perspective of geography, Guangdong Hakka was located in the cluster of south Chinese populations (Figure 2A). Additionally, Guangdong Hakka was observed close to the Southern Han populations (Figure 2B) from the linguistic point of view, while the Guangdong She group was situated in a relatively isolated location with no definite relationship with other Chinese populations from both geographic and linguistic scales (Figure 2).

To further clarify the genetic relationships on a relatively small scale, we performed PCA within Chinese populations (Figure 3). Populations from different regions of China clustered separately, and Guangdong Hakka and Southern Han clustered similarly on the left side. Guangdong She distributed on the middle bottom and clustered with south Chinese populations. However, from the perspective of linguistics, the language relationships between Guangdong She and the surrounding south Chinese populations which are illustrated in Figure 3B were still not distinct.

MDS
MDS plots were conducted based on allele frequencies by Euclidean distance and Manhattan distance, respectively. As shown in Figure 4, each population was represented by a small dot with a different color in the multidimensional space, and the distances between the small dots showed the genetic relationships among the populations in distinct geographic areas or with different language families.
Whether from the perspective of geographical or linguistic scale, the results of MDS analysis by Euclidean distance (Figures 4A,B) had no apparent differences with the principal component analysis (Figures 2A,B). Moreover, in the Manhattan distance-based MDS plots (Figures 4C,D), Guangdong Hakka intertwining with the Southern Han cluster was in accordance with the performances in the MDS conducted by Euclidean distance (Figures 4A,B) and PCA plots (Figure 2). In addition, Guangdong She was located in a relatively isolated place.

Furthermore, to illustrate partial genetic relationships, we collected Guangdong Hakka, Guangdong She, and 83 other Chinese populations to perform MDS analyses with Euclidean distance and Manhattan distance (Figures 4E–H). The MDS results demonstrated that Guangdong Hakka had a close relationship with Southern Han populations, and Guangdong She was situated in a relatively isolated location from the perspectives of geography and linguistics, which were in line with the results of the PCA.

### R_{ST} and Phylogenetic Analyses

The results of the calculations of pairwise $R_{ST}$ and the corresponding $p$ values between Guangdong Hakka and 13 other Chinese Han populations from Southern and Northern Han regions (Kwak et al., 2005; Wu et al., 2011; Zhang et al., 2014;...
Shi et al., 2015; Shu et al., 2015; Bai et al., 2016b; Li et al., 2016; Wang L. et al., 2016; Wang Y. et al., 2016; Zhou et al., 2016, 2018; Nothnagel et al., 2017; Wang H. et al., 2017; Huang et al., 2019; Lang et al., 2019; Sun et al., 2019; Guan et al., 2020) based on 23 Y-STR haplotypes are listed in Supplementary Table 6. The population Guangdong Hakka has been found to be closely related to Guangdong Jieyang Han ($R_{ST} = 0.0028$) and Jiangxi Han ($R_{ST} = 0.0029$). Subsequently, the results of the phylogenetic analysis among Guangdong Hakka and Chinese Han populations are displayed in a phylogenetic tree based on the neighbor-joining method (Figure 5). Guangdong Hakka was first clustered with Guangdong Jieyang Han, followed by Jiangxi Han. Geographically, Jieyang is surrounded by Shanwei, Chaozhou, and Meizhou, whereas the habitations for Guangdong Hakka (Figure 1C) and Jiangxi bordering with Guangdong in the southwest were the dominating Hakka settlements. The phylogenetic tree showed the relationships between Guangdong Hakka and other Han populations in genetics, which were in accordance with the geographical relations. The paternal relationships revealed by Y-STRs indicate that Guangdong Hakka had close relationships with Southern Han, and there were extensive gene flows between Guangdong Hakka and the surrounding Han populations. Our results corroborated the findings of previous studies (Du et al., 2019; Han et al., 2020), which were conducted by using STRs and Y-STRs on the Guangdong Meizhou Hakka population.

From the dimensionality reduction analyses discussed above, the genetic relationships between Guangdong She and other Chinese populations were not clarified. In addition, we employed the Y-STR haplotype profiles of 42 Chinese minorities and Guangdong She to assess the population pairwise genetic distances by AMOVA (Zhu et al., 2005, 2006a,b, 2008; He and Guo, 2013; Shan et al., 2014; Luo et al., 2015, 2019; Ou et al., 2015; Shu et al., 2015; Bai et al., 2016a; Bian et al., 2016; Fu et al., 2016; Gao et al., 2016; Yao et al., 2016; Guo et al., 2017; Ji et al., 2017; Zhao et al., 2017; Cao et al., 2017; Chen et al., 2018; Liu Y. J. et al., 2018; Atif et al., 2019; Liu et al., 2019; Xie et al., 2019; Dezhi et al., 2020; Feng et al., 2020; Song et al., 2020). The pairwise $R_{ST}$ and the corresponding $p$ values between Guangdong She and Chinese minority populations from five linguistically different families are displayed in Supplementary Table 7. No differences were observed between Guangdong She and Guizhou Tujia populations ($R_{ST} = 0.0046$, $p = 0.0574$), while significant genetic differences were observed between Guangdong She and all other Chinese minority populations ($p < 0.05$). The phylogenetic relationships between Guangdong She and 42 Chinese minorities were visualized in the neighbor-joining tree. As shown in
Figure 6, the Tibeto-Burman language-speaking Tibetans and Altaic language-speaking Uighurs and Mongolians clustered together at the upper end, while the language relationships at the bottom, especially for Tai-Kadai, Hmong-Mien, and Tibeto-Burman, were ambiguous. The Guangdong She group clustered with two Tibeto-Burman populations, Guizhou Tujia and Hunan Tujia. However, the Fujian She group congregated with Guizhou Hmong-Mien language-speaking Miao populations.

Prediction of Y-Haplogroups

The above-mentioned phylogenetic analysis hinted that Guangdong She clustered with Tibeto-Burman language-speaking Tujia populations in the same branch. To make further confirmation for the genetic relationships, we employed our in-house database which was composed of 37,754 pieces of Y-SNP/STR data and 109,142 Y-STR (Wang et al., 2015) mainly from East and Southeast Asia to make more precise predictions for 215 Guangdong She males in this study. Finally, 212 out of the 215 genotyped Y-STRs (98.60%) were observed, and a total of six Y-haplogroups were defined in Guangdong She which belong to major clades O2 and O1. The predominantly detected haplogroups were O2-M122 (45.75%), O2a2a1a2-M7 (25.47%), O1a-M119 (10.38%), O1b1a1a-M95 (8.02%), O-M175 (7.08%), and O2a2b1a1-M117 (3.30%), which were determined according to ISOGG, 2019. In addition, a PCA graph was performed among 77 populations (Wen et al., 2004; Gan et al., 2008; Li et al., 2008; Cai et al., 2011; Deng et al., 2013; Fan et al., 2018c; Rowold et al., 2019; Luo et al., 2020; Fan et al., 2021), which were composed of 4,195 individuals in total, including Tai-Kadai, Hmong-Mien, Tibeto-Burman, and Chinese (Southern and Northern Han) populations from Southeast and East Asia, which included three She groups (Fujian, Guangdong Chaoshan, and Guangdong She). From the diagram in Figure 7, the first and second principal components could explain 33.07% of the total variances. Moreover, the populations with different language branches were separated relatively, and the Guangdong She group had a close relationship with the Guangdong Chaoshan She group and

![Figure 7](https://isogg.org/tree/)
clustered with Tibeto-Burman language-speaking populations (including Tujia, Tibetan, and Naxi minorities), while Fujian She located between Hmong-Mien and Han populations, especially Southern Han. The interpopulation comparison demonstrated that (1) different branches of She populations and Fujian and Guangdong She groups may have distinct origins from the perspectives of genetics and linguistics, especially from phylogenetic analyses based on Y-STRs and Y-SNPs, and (2) Guangdong She and Chaoshan She groups have close affinities with Tibeto-Burman language-speaking populations based on the evidence of Y-haplogroups which contained information about the subsequent colonization, differentiations, and migrations overlaid on recent population ranges.

CONCLUSION

In the present study, a total of 475 unrelated Guangdong males (260 Hakka and 215 She) were genotyped by using 23 Y-STRs (including 17 Yfiler and 6 additional Y-STRs) by Promega PowerPlex® Y23 System. For Guangdong Hakka, a total of 176 different alleles were found, with corresponding allelic frequencies ranging from 0.0038 to 0.7423 and GD values that varied from 0.4877 to 0.9671, and the systematic effectiveness for Guangdong Hakka was observed to be sufficient (HD = 0.9994, DC = 0.8885). From the perspectives of geographical and linguistic scales, the phylogenetic analyses indicated that Guangdong Hakka had a close relationship with Southern Han, and there were extensive gene flows between Guangdong Hakka and the surrounding Han populations. For Guangdong She, we identified 155 distinct alleles with a range of allele frequencies from 0.0047 to 0.8605. The GD values for 23 Y-STRs ranged from 0.3277 to 0.9526, and the overall DC and HD were 0.5674 and 0.9939, respectively. The predominant haplogroups of the Guangdong She group were O2-M122 and O2a2a1a2-M7. Moreover, Guangdong She clustered with Tibeto-Burman language-speaking populations (Guizhou Tujia and Hunan Tujia), which demonstrates that the Guangdong She group seems to be one branch of Tibeto-Burman populations and the Huonie dialect of She languages may be a branch of Tibeto-Burman language families.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/Supplementary Material.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Medical Ethics Committee of Hainan Medical University. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

HF contributed to conceptualization, formal analysis, visualization, and writing—original draft preparation. CL and DP took charge of resources and contributed to project administration. HF, LW, and KR were in charge of software. LD, YL, QX, and YZ conducted the investigation. LD, YZ, FW, and ZD performed the validation. CL contributed to data curation. QX, SN, and MJ contributed to writing—review and editing. PQ and S-QW supervised the study. CL, HF, FW, and ZD took charge of funding acquisition. All authors reviewed the manuscript.

FUNDING

This study was supported by grants from the Program of Heyuan for Social Development and Technology Plans (No. 201911810008 and 201911810023), the Program of Hainan Association for Science and Technology Plans to Youth R&D Innovation (QCXM201705), and the National Undergraduate Innovation and Entrepreneurship Training Program (Nos. 201911810008 and 201911810023).

ACKNOWLEDGMENTS

We would like to thank the donors who contributed samples for this study.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.676917/full#supplementary-material

REFERENCES

Adnan, A., Ralf, A., Rakha, A., Kousouri, N., and Kayser, M. (2016). Improving empirical evidence on differentiating closely related men with RM Y-STRs: a comprehensive pedigree study from Pakistan. Forensic Sci. Int. Genet. 25, 45–51. doi: 10.1016/j.fsigen.2016.07.005
Al-Snan, N. R., Messaoudi, S. A., Khubrani, Y. M., Wetton, J. H., Jobling, M. A., and Bakhiet, M. (2020). Geographical structuring and low diversity of paternal lineages in Bahrain shown by analysis of 27 Y-STRs. Mol. Genet. Genom. 295, 1315–1324. doi: 10.1007/s00438-020-01696-4
Atif, A., Kasim, K., Rakha, A., Noor, A., Cheema, A. S., Hadi, S., et al. (2019). Population data of 23 Y STRs from Manchu population of Liaoning Province, Northeast China. Int. J. Legal. Med. 133, 785–788. doi: 10.1007/s00414-018-1957-7
Bai, R., Liu, Y., Ix, X., Shi, M., and Ma, S. (2016a). Genetic polymorphisms of 17 Y chromosomal STRs in She and Manchu ethnic populations from China. Forensic Sci. Int. Genet. 22, e12–e14. doi: 10.1016/j.fsigen.2016.01.011
Bai, R., Liu, Y., Zhang, J., Shi, M., Dong, H., Ma, S., et al. (2016b). Analysis of 27 Y-chromosomal STR haplotypes in a Han population of Henan province,
performance of haplogroup prediction tools. *Forensic Sci. Int. Genet.* 44,102163. doi: 10.1016/j.fsigen.2019.102163

Ji, J., Ren, Z., Zhang, H., Wang, Q., Wang, J., Kong, Z., et al. (2017). Genetic profile of 23 Y chromosomal STR loci in Guizhou Shui population, southwest China. *Forensic Sci. Int. Genet.* 28, e16–e17. doi: 10.1016/j.fsigen.2017.01.010

Juan, M., Xang, W., Yanmei, H., Huiyong, J., Liwei, G., Qian, Z., et al. (2018). Genetic polymorphism of 27 Y-STR Loci among Xianxiang Han population in Henan Province (in Chinese). *Med. Sci.* 54, 452–457. doi: 10.13705/j.issn.1671-6825.2017.09.150

Jun, Y., Li, W., Jing, G., Jiaxin, X., Jinfeng, G., and Baojie, W. (2017). Investigation on the allele frequencies of 41 Y-STR Loci in Han Population in Jilin (in Chinese). *Biol. Chem. Eng.* 6, 103–107.

Khubrani, Y. M., Wetton, J. H., and Jobling, M. A. (2018). Extensive geographical and social structure in the paternal lineages of Saudi Arabia revealed by analysis of 27 Y-STRs. *Forensic Sci. Int. Genet.* 33, 98–105. doi: 10.1016/j.fsigen.2017.11.015

Kumar, S., Stecher, G., and Tamura, K. (2016). MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. *Mol. Biol. Evol.* 33, 1870–1874. doi: 10.1093/molbev/msw054

Kwak, K. D., Jin, H. J., Shin, D. I., Kim, J. M., Roewer, L., Krawczak, M., et al. (2015). Y-chromosomal STR haplotypes and their applications to forensic and population studies in east Asia. *Int. J. Legal. Med.* 119, 195–201. doi: 10.1007/s00414-014-0518-4

Lacerenza, D., Aneli, S., Di Gaetano, C., Critelli, R., Piazza, A., Matullo, G., et al. (2017). Investigation of extended Y chromosome STR haplotypes in Sardinia. *Int. J. Legal. Med.* 131, 195–200. doi: 10.1007/s00414-016-1424-z

Letunic, I., and Bork, P. (2019). Interactive Tree Of Life (iTOL) v4: recent updates and enhancements. *Evol. Biol.* 42, e13–e20. doi: 10.1007/s00439-018-0566-5

Liu, Y., Jin, X., Guo, Y., Zhang, X., Zhu, W., Zhang, W., et al. (2020). Haplotypic analysis of the polymorphic 40 Y-STR markers in Chinese populations. *Forensic Sci. Int. Genet.* 33, 98–105. doi: 10.1016/j.fsigen.2017.11.015

Liu, Y. J., Guo, L. H., Li, J., Yue, J. T., and Shi, M. S. (2018). Genetic polymorphism of 27 Y-STR Loci in Dongxiang population of Gansu province. *Fa Yi Xue Za Zhi* 34, 270–275. doi: 10.12116/j.issn.1004-5619.2018.03.011

Liu, J. N., and Wang, C. C. (2019). A new approach to the origin of the she in the perspective of molecular anthropology. *Philos. Soc. Sci.* 45, 98–107.

Luo, H., Song, F., Zhang, L., and Hou, Y. (2015). Genetic polymorphism of 23 Y-STR loci in the Zhuang minority population in Guangxi of China. *Int. J. Legal. Med.* 129, 737–738. doi: 10.1007/s00414-015-1175-5

Luo, X. L. (1989). *The Orign of Hakka*. Beijing: The Chinese Overseas Publishing House.

Luo, X. Q., Du, P. X., Wang, L. X., Zhou, B. Y., Li, Y. C., Zheng, H. X., et al. (2020). Uniparental Genetic analyses reveal the major origin of Fujian Tanka from ancient indigenous Daic populations. *Hum. Biol.* 91, 257–277. doi: 10.13110/humanbiology.91.4.05

Luo, Y., Wu, Y., Qian, E., Wang, Q., Zhang, H., et al. (2019). Population genetic analysis of 36 Y-chromosomal STReS Yields comprehensive insights into the forensic features and phylogenetic relationship of Chinese Tai-Kadai-speaking Bouyei. *PloS One* 14:e0224601. doi: 10.1371/journal.pone.0224601

Mozina, M., Fan, G., Guo, F., He, Y., Hou, Y., Hu, S., et al. (2017). Revisiting the male genetic landscape of China: a multi-center study of almost 38,000 Y-STR haplotypes. *Hum. Genom.* 136, 485–497. doi: 10.1002/sphg.1759-x

Oppenheimer, S. (2012). Out-of-Africa, the peopling of continents and islands: tracing uniparental gene trees across the map. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 367, 770–784. doi: 10.1098/rstb.2011.0306

Ou, X., Wang, Y., Liu, C., Yang, D., Zhang, C., Deng, S., et al. (2015). Haploptype analysis of the polymorphic 40 Y-STR markers in Chinese populations. *Forensic Sci. Int. Genet.* 19, 255–262. doi: 10.1016/j.fsigen.2015.08.007

Pickrill, I., Muller, E., Zaher, W., Dunkelmann, B., Cemper-Kieslich, J., Keindl, G., et al. (2016). Yfiler(R) Plus amplification kit validation and calculation of forensic parameters for two Austrian populations. *Forensic Sci. Int. Genet.* 21, 90–94. doi: 10.1016/j.fsigen.2015.12.014

Prinz, M., Boll, K., Baum, H., and Shaler, B. (1997). Multiplexing of Y chromosome specific STReS and performance for mixed samples. *Forensic Sci. Int.* 85, 209–218. doi: 10.1016/s0378-7758(96)02096-8

Qiang, L., Yue, X., Wei, Z., Dian, Z., Baowen, C., and Faming, Z. (2017). Genetic Polymorphism Analysis of 27 Y-STR Loci in the Hui Population from Weishan, Yunnan Province (in Chinese). *J. Kunming Med. Univ.* 38, 25–29.

Rowe, L., Andersen, M. M., Ballantine, J., Butler, J. M., Caliebe, A., Corach, D., et al. (2020). DNA commission of the International Society of Forensic Genetics (ISFG): recommendations on the interpretation of Y-STR results in forensic analysis. *Forensic Sci. Int. Genet.* 48:102308. doi: 10.1016/j.fsigen.2020.10.2308

Rowold, D. J., Gayden, T., Luis, J. R., Alfonso-Sanchez, M. A., Garcia-Bertrand, R., and Herrerera, R. J. (2019). Investigating the genetic diversity and affinities of historical populations of Tibet. *Gene* 682, 81–91. doi: 10.1016/j.gene.2018.09.043

Saitou, N., and Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* 4, 406–425. doi: 10.1093/oxfordjournals.molbev.a040454

Santos Stange, V., Silva Dos Reis, R., Mariano Garcia de Souza, R., Lima, L. M., Machado, C. M., and Mazando, S. (2020). Genetic analysis of the polymorphic 40 Y-STR markers in Brazilian populations of Espirito Santo for 27 Y-STRs. *Forensic Sci. Int.* 367, 770–784. doi: 10.1016/j.legalmed.2019.101660

Shonhai, M., Nhiwatiwa, T., Nangammbi, T., and Mazando, S. (2020). Genetic analysis of the polymorphic 40 Y-STR markers in Brazilian populations of Espirito Santo for 27 Y-STRs. *Forensic Sci. Int.* 367, 770–784. doi: 10.1016/j.legalmed.2019.101660

Shi, M., Liu, Y., Zhang, J., Bai, R., Lv, X., and Ma, S. (2015). Analysis of 24 Y-chromosomal STR haplotypes in a Chinese Han population sample from Henan Province, Central China. *Forensic Sci. Int. Genet.* 17, 83–86. doi: 10.1016/j.fsigen.2015.04.001

Shi, Z. F. (2013). An analysis of the origin and development of she culture. *J. Mudanjiang Coll. Educ.* 144, 123–124.
Shu, L., Li, L., Yu, G., Yu, B., Liu, Y., Li, S., et al. (2015). Genetic analysis of 17 Y-STR loci in Han, Dong, Miao and Tuja populations from Hunan province, central-southern China. Forensic Sci. Int. Genet. 19, 250–251. doi: 10.1016/j.fsigen.2015.07.007

Song, F., Xie, M., Xie, B., Wang, S., Liao, M., and Luo, H. (2020). Genetic diversity and phylogenetic analysis of 29 Y-STR loci in the Tibetban population from Sichuan province. Southwest Chin. Int. J. Legal. Med. 134, 513–516. doi: 10.1007/s00414-019-02043-y

Spólnicka, M., Dabrowska, J., Szablowska-Gnap, E., Paleczka, A., Jabłonska, M., Zbiec-Piekarska, R., et al. (2017). Intra- and inter-population analysis of haplotype diversity in Yfiler(R) Plus system using a wide set of representative data from Polish population. Forensic Sci. Int. Genet. 28, e22–e25. doi: 10.1016/j.fsigen.2017.01.014

Sun, H., Su, K., Fan, C., Long, F., Liu, Y., Sun, J., et al. (2019). Y-STRs’ genetic profiling of 1953 individuals from two Chinese Han populations (Guizhou and Shanxi). Forensic Sci. Int. Genet. 38, e8–e10. doi: 10.1016/j.fsigen.2018.10.011

SWGDAM (2010). Scientific Working Group on DNA Analysis Interpretation Guidelines for Autosomal STR Typing by Forensic DNA Testing Laboratories.

Tao, R., Jin, M., Ji, G., Zhang, J., Zhang, J., and Tai, Y. (2019). Inner and inter population analysis of genetic polymorphisms of 27 Yfiler((R)) Plus loci in the Daur and Mongolian ethnic minorities from Hulunbuir of Inner Mongolia Autonomous Region, China. Forensic Sci. Int. Genet. 40, e252–e255. doi: 10.1016/j.fsigen.2019.02.003

Tao, R., Wang, S., Zhang, J., Zhang, J., Yang, Z., et al. (2019b). Genetic characterization of 27 Y-STR loci analyzed in the Nantong Han population residing along the Yangtze Basin. Forensic Sci. Int. Genet. 39, e10–e13. doi: 10.1016/j.fsigen.2018.11.015

Wang, C. C., Wang, L. X., Shrestha, R., Wen, S., Zhang, M., Tong, X., et al. (2020). Genetic structure and forensic characterisation of 36 Y-chromosomal STR loci in Hmong-Mien-speaking Miao population. Ann. Hum. Biol. 47, 541–548. doi: 10.1080/03014460.2020.1788159

Wang, X., Li, Y., and Fan, H. (2019). The associations between screen time-based sedentary behavior and depression: a systematic review and meta-analysis. BMC Public Health 19:1524. doi: 10.1186/s12889-019-7904-9

Wen, B., Li, H., Lu, D., Song, X., Zhang, F., He, Y., et al. (2004). Genetic evidence supports demic diffusion of Han culture. Nature 431, 302–305. doi: 10.1038/nature02878

Westen, A. A., Kraaijenbrink, T., Clarisse, L., Grol, L. J., Willems, P., Zuniga, S. B., et al. (2015). Analysis of 36 Y-STR marker units including a concordance study among 2085 Dutch males. Forensic Sci. Int. Genet. 14, 174–181. doi: 10.1016/j.fsigen.2014.10.012

Wu, W., Pan, L., Hao, H., Zheng, X., Lin, J., and Lu, D. (2011). Population genetics of 17 Y-STR loci in a large Chinese Han population from Zhejiang Province, Eastern China. Forensic Sci. Int. Genet. 5, e11–e13. doi: 10.1016/j.fsigen.2009.12.005

Xie, M., Song, F., Li, J., Lang, M., Luo, H., Wang, Z., et al. (2019). Genetic substructure and forensic characteristics of Chinese Hui populations using 157 Y-SNPs and 27 Y-STRs. Forensic Sci. Int. Genet. 41, 11–18. doi: 10.1016/j.fsigen.2019.03.022

Yau, L., Tian, M., Chuanzhong, X., Yuebo, L., and Meisen, S. (2019). Genetic polymorphisms of 27 Y-STR loci in Tuja and Hui population and the cluster analysis of 13 ethnic groups (in Chinese). Basic Clin. Med. 39, 314–320.

Yau, L., Jun-tao, Z., Shaoshing, S., Hai, L., Lihong, G., Yi, Z., et al. (2014). Genetic polymorphisms of 27-STR loci in Henna Han population. Forensic Sci. Technol. 4, 18–20.

Yau, L., Lin, G., Juntao, Y., Jin, L., Meisen, S., and Yuebo, L. (2020). Genetic polymorphisms of 27 Y-STR in 3 Chinese populations and the cluster analysis of 13 ethnic groups (in Chinese). Lab. Med. Clin. 5, 577–585. doi: 10.3969/j.issn.1672-9455.2020.05.001

Yao, H. B., Wang, C. C., Tao, X., Shang, L., Wen, S. Q., Zhuo, B., et al. (2016). Genetic evidence for an East Asian origin of Chinese Muslim populations Dongxiang and Hui. Sci. Rep. 6:38656. doi: 10.1038/srep38656

Yin, C., Su, K., He, Z., Zhai, D., Guo, K., Chen, X., et al. (2020). Genetic Reconstruction and Forensic Analysis of Chinese Shandong and Yunnan Han Populations by Co-Analyzing Y Chromosome STRs and SNPs. Genes 11:743. doi: 10.3390/genes11070743

Zayed, T., Adam, A., Alghafiri, R., and Irutin, R. (2020). Study of 27 Y-STR markers in United Arab Emirates population. Forensic Sci. Intern. Rep. 2:100057.

Zgonjanin, D., Alghafiri, R., Antov, M., Stojilkovic, G., Petkovic, S., Vukovic, R., et al. (2017). Genetic characterization of 27 Y-STR with the Yfiler((R)) Plus kit in the population of Serbia. Forensic Sci. Int. Genet. 31, e48–e49. doi: 10.1016/j.fsigen.2017.07.013

Zhahbin, M., Sarkytbaveya, A., Tazhiloglu, I., Yerezhепов, D., Li, S., Akilzhahanov, R., et al. (2019). Development of the Kazakhstan Y-chromosome haplotype reference database: analysis of 27 Y-STR in Kazakh population. Int. J. Legal. Med. 133, 1029–1032. doi: 10.1007/s00414-018-1589-8

Zhong, J., Tao, R., Zhong, J., Sun, D., Qiao, L., Shan, S., et al. (2019). Genetic polymorphisms of 27 Y-STR loci in the Dezhou Han population from Shandong province, Eastern China. Forensic Sci. Int. Genet. 39, e26–e28. doi: 10.1016/j.fsigen.2018.11.021

Zhong, J., Wang, J., Liu, Y., Shi, M., Bai, R., and Ma, S. (2017). Haplotyping data for 27 Y-chromosomal STR loci in the Chaoshan Han population, South China. Forensic Sci. Int. Genet. 31, e54–e56. doi: 10.1016/j.fsigen.2017.08.003

Zhang, S., Tian, H., Wang, Z., Zhao, S., Hu, Z., Li, C., et al. (2014). Development of a new 26plex Y-STR typing system for forensic application. Forensic Sci. Int. Genet. 13, 112–120. doi: 10.1016/j.fsigen.2014.06.015

Zhao, Q., Bian, Y., Zhang, S., Zhu, R., Zhou, W., Gao, Y., et al. (2017). Population genetics study using 26 Y-chromosomal STR loci in the Hui ethnic group in China. Forensic Sci. Int. Genet. 28, e26–e27. doi: 10.1016/j.fsigen.2017.01.018

Zheng, L., Li, Y., Lu, S., Bao, J., Wang, Y., Zhang, X., et al. (2013). Physical characteristics of Chinese Hakka. Sci. China Life Sci. 56, 541–551. doi: 10.1007/s11427-013-4471-7

Zhang, J. K. (2019). The review and prospect of the study in Hakka historical origin. Local Cult. Res. 02, 101–122.

Zhou, H., Ren, Z., Zhang, H., Wang, J., and Huang, J. (2016). Genetic profile of 17 Y chromosome STRs in the Guizhou Han population of southwestern China. Forensic Sci. Int. Genet. 25, e6–e7. doi: 10.1016/j.fsigen.2016.05.010
Zhou, Y., Shao, C., Li, L., Zhang, Y., Liu, B., Yang, Q., et al. (2018). Genetic analysis of 29 Y-STR loci in the Chinese Han population from Shanghai. Forensic Sci. Int. Genet. 32, e1–e4. doi: 10.1016/j.fsigen.2017.11.003

Zhu, B., Deng, Y., Zhang, F., Wei, W., Chen, L., Zhao, J., et al. (2006a). Genetic analysis for Y chromosome short tandem repeat haplotypes of Chinese Han population residing in the Ningxia province of China. J. Forensic Sci. 51, 1417–1420. doi: 10.1111/j.1556-4029.2006.00282.x

Zhu, B., Liu, S., Ci, D., Huang, J., Wang, Y., Chen, L., et al. (2006b). Population genetics for Y-chromosomal STRs haplotypes of Chinese Tibetan ethnic minority group in Tibet. Forensic Sci. Int. 161, 78–83. doi: 10.1016/j.forsciint.2005.09.003

Zhu, B., Li, X., Wang, Z., Wu, H., He, Y., Zhao, J., et al. (2005). Y-STRs haplotypes of Chinese Mongol ethnic group using Y-PLEX 12. Forensic Sci. Int. 153, 260–263. doi: 10.1016/j.forsciint.2004.11.002

Zhu, B., Wu, Y., Shen, C., Yang, T., Deng, Y., Xun, X., et al. (2008). Genetic analysis of 17 Y-chromosomal STRs haplotypes of Chinese Tibetan ethnic group residing in Qinghai province of China. Forensic Sci. Int. 175, 238–243. doi: 10.1016/j.forsciint.2007.06.012

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The handling editor AA declared a past collaboration with the author SN.