9-21-2022

Y-STR Haplotypic Polymorphisms for the Hakka Population in West China and Its Phylogenic Comparison with Other Chinese Populations

Meng-Nan Liu
Shaoxing University

Chang-Xiu Peng
Shaoxing University

Dan-Lu Song
3Ningbo Health Gene Technologies Co. Ltd.

Hai-Ying Jin
Ningbo Health Gene Technologies Co. Ltd.

Xing-Kai Zheng
Ningbo Health Gene Technologies Co. Ltd.

See next page for additional authors

Y-STR Haplotypic Polymorphisms for the Hakka Population in West China and Its Phylogenic Comparison with Other Chinese Populations

Meng-Nan Liu,[1] Chang-Xiu Peng,[2#] Dan-Lu Song,[3] Hai-Ying Jin,[3] Xing-Kai Zheng,[3] and Guang-Yao Fan[4*]

[1]College of Medicine, Shaoxing University, Shaoxing 312000, China.

[2]College of Life Sciences, Shaoxing University, Shaoxing 312000, China.

[3]Ningbo Health Gene Technologies Co. Ltd., Ningbo 315040, China.

[4]Forensic Center, College of Medicine, Shaoxing University, Shaoxing 312000, China.

[#]The author contributed equally to this work and should be considered co-first author.

*Correspondence to: Guang-Yao Fan, No. 508 Huancheng West Road, Shaoxing 312000, China. E-mail: fanyoyo1983@163.com.

Short Title: Y-STR Analysis for the Hakka Population in West China

KEY WORDS: Y-CHROMOSOMAL STR, POPULATION GENETICS, HAPLOTYPE DIVERSITY, PHYLOGENETIC RECONSTRUCTION, DONGSHAN HAKKA.

Abstract

The Hakkas, undergone a series of great migrations, are usually identified with people who speak the Hakka language or share at least same Hakka ancestry. As the largest Hakka dialect island in West China, the Dongshan region was closely linked with the great migration wave of Hakka. However, the paternal genetic profiles of Dongshan Hakka have never been revealed. In the present study, 41 Y-chromosomal short tandem repeat (Y-STR) loci included in the SureID® PathFinder Plus Kit were analyzed in 353 unrelated male individuals (171 Hakka and 182 Han) of Sichuan Province, China. By analyzing 166 different haplotypes among Dongshan Hakkas and 176 different haplotypes among Sichuan Han males, haplotype diversity (HD) of the Hakka population was calculated as 0.9997 with a discrimination capacity (DC) of 0.9708. HD and DC were 0.9996 and 0.9670 for the Sichuan Han population, respectively. Most of the Y-STR loci were highly informative in both populations except DYS645. The genetic relationships were evaluated by comparing the Hakka population with 11 other groups that are relevant to the migration routes of Hakkas. The results of the MDS plot and phylogenetic tree indicate that the Dongshan Hakka population was closely related to Han nationalities from Anhui, Jiangxi, and Fujian Provinces.

Hakka is one of the seven major Han Chinese subgroups (Du et al. 2019), whose culture is regarded as the living fossil of ancient Han culture (Tan 2008). They not only kept their own ethnic culture tenaciously but also integrated with the native culture along their migration routes. In general, the Hakka have experienced five great migrations in the trek over more than a thousand years. During the fourth great migration wave (1671-1776 A.D.) numerous Hakka people migrated from Central and Southeastern China to Sichuan and reached the western foot of the Longquan Mountains (Dongshan) (Supplementary Figure S1). This historical event was also called the immigration from Hu-Guang to Sichuan Province (Li et al. 2012), which was authenticated by many genealogical and linguistic evidence. Until now, Dongshan is the largest Hakka dialect island in West China, and attracted scholarly interests (Su 2017). The linguistic characteristics of the Hakka dialect are significantly different from Sichuanese Mandarin and the other families of Sino-Tibetan languages (Hashimoto 1973). However, the paternal genetic profiles of Dongshan Hakka have never been revealed.

Y-chromosome is widely regarded as one of the most important tools for studies of forensic and human population genetics (Bian et al. 2016). Genetic markers in the non-recombined part of Y-chromosome have unique advantages in tracking male offspring(Balanovsky 2017). With high mutability, Y-chromosome short tandem repeats (Y-STRs) were widely used to study patrilineal diversity in various populations and elucidate human population history. However, Y-STR has not yet been investigated for a comprehensive analysis of the Dongshan Hakka population.

Materials and Methods

Sample Collection

In this study, buccal and blood samples on FTA® storage cards were collected from 353 unrelated healthy individuals of Dongshan Hakka (n = 171) and Sichuan Han populations (n = 182) (Supplementary Table S1). All the Dongshan Hakka subjects speak Hakka dialect. The sampling sites of Han populations are located inside the Sichuan Basin in three different directions (South, East, and North) (Supplementary Figure S1). We confirmed that the subjects of Hakka and Han had lived in their regions for at least three generations. This study was approved by the Ethics Committee of Medical College, Shaoxing University. All participants signed the informed consent forms according to the Declaration of Helsinki (Carlson et al. 2004).

PCR Amplification and Y-STR Genotyping

Following the technical instructions in previous research (Fan et al. 2021), 41 Y-STR loci and 3 Y-InDel markers were co-amplified by the SureID® PathFinder Plus Kit. PCR was performed on GeneAmp® PCR System 9700 Thermal Cycler (Applied Biosystems, CA, USA). PCR products were separated and detected on ABI 3500 Genetic Analyzer (Applied Biosystems, USA). The analysis of genotyping was conducted by GeneMapper® *ID-X* (Applied Biosystems, USA).

Haplotypic Polymorphisms Analyses

Haplotype and allele frequency of Y-STR loci were calculated by direct counting. Genetic diversity (GD) and haplotype diversity (HD) were measured using Nei's method (Nei and

Tajima 1981). The discrimination capacity (DC) was represented as the ratio between total distinct haplotypes and the total number of haplotypes.

Population Substructure Reconstruction

Furthermore, genetic relationships were compared among Dongshan Hakka, and Sichuan Han from four directions, and other seven reference populations associated with Hakka's migration (Supplementary Table S1). The R_{ST} genetic distances were assessed by analysis of molecular variance (AMOVA) and visualized in multidimensional scaling (MDS) plot using the online calculation tool of YHRD (https://yhrd.org/amova). The Y-STR haplotype data are available in the YHRD under accession number YA005897 for the Sichuan Hakka population and YA004694-2 for the Sichuan Han population. A neighbor-joining phylogenetic tree was constructed using Phylogeny Inference Package (PHYLIP) v3.6.95 (Reynolds et al. 1983) and visualized by Evolview v3 (Subramanian et al. 2019). The description of population substructures was also determined using linear discriminant analysis (LDA, aka Fisher discriminant analysis) (Diaz-Vico and Dorronsoro 2020). Based on 33 Y-STR loci the LDA plot was created via the open-source script of R 4.1.3 to view the linear discriminant of the model and visualize how well it separated the four different ethnic groups in Sichuan Province. The multi-copy loci were excluded in the LDA.

Paternal Haplogroup Assignment

To allocate Y-SNP haplogroups to the respective individuals, the k-nearest neighbor (kNN) prediction model was utilized (Altman 1992). As a common tool for machine-learning, its

good effects on prediction were approved in many previous studies (Liong and Foo 2013; Yin et al. 2022). The kNN depends on a large amount of training data. In order to improve the prediction performance of the kNN model, Y-STR haplotypes of 3248 samples from Han populations and their corresponding Y haplogroups were collected to form the training and testing dataset (Lang et al. 2019; Song et al. 2019; Yin et al. 2020; Yin et al. 2022; Zhang et al 2020). The program was implemented by the open-source script of R 4.1.3 based on 23 shared Y-STR loci (Lang et al. 2019; Song et al. 2019; Yin et al. 2020; Yin et al. 2022; Zhang et al 2020). There was no intermediate allele variants or allele sizes limitation. Meanwhile, copy number variation (CNV) of Y-STR alleles and multi-copy loci were excluded in machine-learning (ML) development. Before the Y-haplogroup assignment, all the rest samples were classified into eight consolidated haplogroup branches (C2b1, D1a1, N1a1, N1a2, O1a1, O2a1, O2a2, and R1a1). The binary genetic markers which were used to defined the haplogroup branches are indicated in Supplementary Table S2, respectively. The ratio of the training set to testing set was set as 7:3. A confusion matrix was generated to present the performance for the prediction. The sensitivity and specificity for each predicted haplogroup were also calculated. The program was implemented by the open-source script of R 4.1.3 based on 23 shared Y-STR loci.

Network Construction

Network 10.2. was used to generate the classic median-joining network (Bandelt, et al. 1999). Only complete haplotypes for the 33 Y-STR loci in the SureID® PathFinder Plus Kit were used for network construction, while the multi-copy loci of DYS385, DYF387S1, DYS527,

and DYF404S1 were excluded ahead. Meanwhile, considering the Y-STR variation across the haplogroups in the four studied populations, the weights assigned were specific with a five-fold range for each haplogroup (Qamar, et al. 2002; Fan et al. 2022).

This work follows the updated guidelines for the publication of population data requested by the journal (Carracedo et al. 2013; Carracedo et al. 2014).

Results and Discussion

Genetic Diversity

The haplotype data of 41 Y-STR loci in the populations of Dongshan Hakka and Sichuan Han are shown in Supplementary Table S3 and Supplementary Table S4, respectively. Allelic frequency and GD values are listed in Supplementary Table S5. GD values at most of the 41 Y-STR loci were greater than 0.5 (Supplementary Figure S2). The highest GD value (0.8873) at DYS449 and the lowest one (0.0645) at DYS645 were observed both in the Han population. The allelic combination distribution of the four multi-copy loci is presented in Supplementary Table S6. In addition, copy number variants were only found in these multiple-copy loci (Supplementary Table S7).

Comparison with Other Populations

Pairwise R_{ST} genetic distances among 12 studied populations, and the associated p values, are presented in Supplementary Table S8. When we focus on the Hakka population in this result, the genetic distance between Hakka and East Sichuan Han is the smallest (-0.0007), while the genetic distance between Hakka and Shaanxi Han is the largest (0.0220). Geographically, the

sampling sites of Sichuan Han (South, East, and North) are adjacent to the entrances of natural barriers of the Sichuan Basin, where the Hakka immigrants were hard to detour (Supplementary Figure S1). This special terrain made it possible to associate the sampling sites with the potential migration routes. The MDS results suggested that Dongshan Hakka was relatively far from the West Sichuan Han (Figure 1), although they are geographically much closer than any other group. Among the four studied ethnic groups of Sichuan, Dongshan Hakka showed a closer relationship with East Sichuan Han, even though all the Sichuan Han share the same dialect. This finding can be attributed to the important waterway of the eastern route. Numerous Hakka ancestors are believed to enter the eastern part of the Sichuan Basin along the Yangtze River and settle there in the early stage (Zhan 2013). Sichuan is geographically close to the ethnolinguistically diverse provinces of Chongqing, Guizhou, Yunnan, Hunan, and Guangxi (Wang et al. 2021). Despite this, among all the reference ethnic groups, East Sichuan Han mapped relatively close to Anhui Han and Jiangxi Han in the MDS plot. It was also noteworthy that Dongshan Hakka, East Sichuan Han, and Hunan Miao were not far from each other. A former ethnolinguistic investigation, which suggested that the Hakka dialect had intermixed with the Hmong-Mien language to some extent, may account for the affinity between the two ethnic groups (Li 2007). However, compared to the genetic affinities between Miao and Han populations from Anhui, Jiangxi, and Sichuan (Figure 1), Hakka was relatively farther from Miao. Consistent with the previous studies (Shi et al. 2005; Wang 1994), the Hmong-Mien populations were clustered closely with Han populations, which reflects the recorded history of admixture. Besides, linear discriminant analysis is a useful method to classify a response variable into two or more

classes. Based on 33 single-allele Y-STR loci, we assessed the performance of Y-STR haplotypes in classifying different ethnic groups. As shown in Figure 2, LDA was able to illustrate the mapping space for each of the haplotypes. The clearer substructures were observed in the North Sichuan Han and East Sichuan Han, while the South Sichuan Han was intermixed with Hakka. The findings were inconsistent with a previous study which suggested the homogeneity of the Sichuan Han population in many microareas (Fan et al. 2017). This controversy may attribute to the limited number of Y-STR loci utilized in that study.

Phylogenetic Reconstruction

To further illustrate the phylogenetic relationships among the Dongshan Hakka, Sichuan Han, and other reference populations, a neighbor-joining tree was constructed based on the matrix of R_{ST} genetic distances. As shown in Figure 3, Dongshan Hakka was first clustered with Anhui Han, followed by Han populations from Jiangxi and Fujian (Longyan). It may be associated with the long-term fusion between Hakka and native Han people after the second great migration wave. These findings were consistent with another research which suggested Hubei, Canton, and Jiangxi are all the possible origins of Hakka (Li et al. 2003). Nevertheless, the close genetic relationship was not observed between Guangdong Han and Dongshan Hakka as expected in the present study. Considering the former reports which revealed a close affinity between Meizhou Hakka and Guangdong Han (Du et al. 2019; Han et al. 2019), one of the reasonable explanations is that one or more branches of Hakka had separated before their integration with Guangdong Han. There are millions of people who

migrated toward Sichuan during the fourth great migration wave which is a famous demographic event in Chinese history. That duration might span multiple generations. To verify the main route of the immigration from Hu-Guang to Sichuan Province, more investigations of the relevant populations are still needed, especially for the Han nationalities of Hunan and Hubei. In addition, the results of the phylogenetic tree had shown that Guangdong Han, as well as Han populations from East and South Sichuan, gathered in one clade (Figure 3). It hints that numerous Han people might move from Guangdong to Sichuan analog to the westward migration of Hakka. Interestingly, the Dongshan Hakka was phylogenetically close to Anhui rather than Guangdong and Fujian which are thought to be the origin places of the Hakka's fourth great migration. A former study revealed the differentiated split histories and founder effects for the clans of Cantonese, Hakka, and Minnan Chaoshanese (Lan et al. 2020). The clear separation of Hakka and Southeastern Han Chinese in the phylogenetic tree may be corresponding to cultural-linguistic segregation and longtime conflicts (Hakka-Punti Wars) within the clans before immigration waves. Of course, this doesn't mean that the genetic affinity between Cantonese and Guangdong Hakka was nonexistent. On the contrary, the national fusion may be more significant, especially in modern Chinese history. As expected, another phylogenetic analysis via Y-STR shown us that Guangdong Hakka has a close relationship with Southern Han regarding geographical and linguistic scales (Luo et al. 2021). In contrast, Dongshan Hakka in our study seems to retain a relatively isolated genetic background.

ML-based Haplogroup Prediction

As the Y-STR haplotypes are genetically related to Y-SNP haplogroups, many haplogroup prediction models were developed to predict Y-SNP haplogroups in Han Chinese by Y-STR haplotypes (Yin et al. 2022). The kNN model was proved to be one of the optimal approaches for this purpose (Yin et al. 2022). the efficiency of kNN model in predicting eight detailed haplogroups (C2b1, D1a1, N1a1, N1a2, O1a1, O2a1, O2a2, and R1a1) was estimated before the haplogroup prediction. The confusion matrix showed that the Y-SNP haplogroups of most of the samples were predicted correctly (Supplementary Table S9). Accuracy scores of the kNN predictor are summarized in Supplementary Table S10 (overall average accuracy: 97.86%). The specificity for the eight haplogroups prediction had good performances, C2b1 (99.87%), D1a1 (99.97%), N1a1(100.00%), N1a2(99.91%), O1a1 (99.87%), O2a1 (98.28%), O2a2 (98.56%) and R1a1 (100.00%). In Figure 4, all the eight haplogroups existed among the Hakka and South Sichuan Han populations, while the N1a2 in North Sichuan Han and N1a1 East Sichuan Han were lost, respectively. The first two dominant haplogroups were O2a2 and O2a1 in all four populations. A previous study had showed that the O2-M122 lineages were probably dominant in the Hmong-Mien populations such as Yunan Miao 43.8% (21/48) and Hunan Miao 45.7% (48/105) (Wen et al. 2005). The frequency distribution of O2 haplogroups in the four subpopulations of Sichuan was very similar to the reported Miao. This was consistent with the findings of Figure 1 in which a closed relationship between Hunan Han and Sichuan Han was presented. According to recorded history, the Hmong-Mien populations had undergone an admixture with Han populations (Shi et al. 2005).

Network Analysis

The median-joining Y-STR networks for O2a1 and O2a2 were calculated by Network 10.2. (Figure 5). Two and three distinctive clusters were revealed in the networks of O2a1 and O2a2, respectively. The Dongshan Hakka males nearly lay in all the branches of the two networks. Neither the Sichuan Han nor the Dongshan Hakka could form a distinct cluster. It may attribute to their closely related haplotypes with each other. Notably, South Sichuan Han appeared in each of the branches accompanied by Hakka. This complex pattern may indicate that South Sichuan Han has a more significant impact on the paternal lineages of Dongshan Hakka. This was consistent with the findings in LDA.

Conclusion

In conclusion, these haplotypic data on the Dongshan Hakka and Sichuan Han populations have the potential for studying genealogy and uncovering the male-specific migration history. Dongshan Hakka population was phylogenetically close to Han nationalities from Anhui, Jiangxi, and Fujian Provinces. Moreover, the dominant haplogroups were O2a2 and O2a1 in both Hakka and Sichuan Han populations. However, it is still a challenging task to exact depict the migration routes of Hakka. Our study is expected to have a positive influence on the investigation of Hakka's migration history over the coming years.

Conflict of Interest

The authors declare that they have no conflict of interest.

Acknowledgments

The authors are grateful to the voluntary participants in this study. This work was supported by the National Social Science Fund of China (Grant No. 21BMZ006).

Received 8 December 2021; accepted for publication 28 May 2022.

Literature Cited

Altman, N. S. 1992. An introduction to kernel and nearest-neighbor nonparametric regression. *Am. Stat.* 46:175–185.

Balanovsky, O. 2017. Toward a consensus on SNP and STR mutation rates on the human Y-chromosome. *Hum. Genet.* 136:575–590.

Bandelt, H. J., P. Forster, and A. Rohl. 1999. Median-joining networks for inferring intraspecific phylogenies. *Mol. Biol. Evol.* 16:37–48.

Bian, Y., S. Zhang, W. Zhou et al. 2016. Analysis of genetic admixture in Uyghur using the 26 Y-STR loci system. *Sci. Rep.* 6:19998.

Carlson, R. V., K. M. Boyd, and D. J. Webb. 2004. The revision of the Declaration of Helsinki: Past, present and future. *Br. J. Clin. Pharmacol.* 57:695–713.

Carracedo, A., J. M. Butler, L. Gusmao et al. 2013. New guidelines for the publication of genetic population data. *Forensic Sci. Int. Genet.* 7:217–220.

Carracedo, A., J. M. Butler, L. Gusmao et al. 2014. Update of the guidelines for the publication of genetic population data. *Forensic Sci. Int. Genet.* 10:A1–A2.

Díaz-Vico, D., and J. R. Dorronsoro. 2020. Deep least squares Fisher discriminant analysis. *IEEE Trans. Neural Netw. Learn. Syst.* 31:2,752–2,763.

Du, W., W. Wu, Z. Wu et al. 2018. Genetic polymorphisms of 32 Y-STR loci in Meizhou Hakka population. *Int. J. Legal Med.* 133:465–466.

Fan, G., W. Li, Y. Ye et al. 2017. Haplotype diversity of 17 Y-chromosome STR loci in Han population from different areas of Sichuan Province, Southwest China. *Leg. Med. (Tokyo)* 26:73–75.

Fan, G., L. Pan, P. Tang et al. 2021. Technical note: Developmental validation of a novel 41-plex Y-STR system for the direct amplification of reference samples. *Int. J. Legal Med.* 135:409–419.

Fan, G. Y., D. L. Song, H. Y. Jin et al. 2022. Gene flow and phylogenetic analyses of paternal lineages in the Yi-Luo valley using Y-STR genetic markers. *Ann. Hum. Biol.* 48:627–634.

Han, X., A. Shen, T. Yao et al. 2021. Genetic diversity of 17 autosomal STR loci in Meizhou Hakka population. *Int. J. Legal Med.* 135:443–444.

Hashimoto, M. J. 1973. *The Hakka Dialect: A Linguistic Study of its Phonology, Syntax, and Lexicon.* Cambridge: Cambridge University Press.

Lan, A., K. Kang, S. Tang et al. 2020. Fine-scale population structure and demographic history of Han Chinese inferred from haplotype network of 111,000 genomes. Preprint, https://www.biorxiv.org/content/10.1101/2020.07.03.166413v2, doi: 10.1101/2020.07.03.166413.

Lang, M., H. Liu, F. Song 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.

Li, H. 2007. Abscondence of Min-Yue ethnic group revealed by molecular anthropology. *J. Guangxi Univ. Natl. (Philos. Soc. Sci. Ed.)* 29:42–47.

Li, H., W.-Y. Pan, B. Wen et al. 2003. Origin of Hakka and Hakkanese: A genetics analysis. *Yi Chuan Xue Bao* 30:873–880.

Li, J., A. Deng, Y. Chen et al. 2012. Current situation and prospect of the study on Sichuan

Hakka folk houses. *Appl. Mech. Mater.* 174–177:1,639–1,644.

Liong, C. Y., and S. F. Foo. 2013. Comparison of linear discriminant analysis and logistic regression for data classification. *AIP Conf. Proc.* 1522:1,159–1,165.

Luo, C., L. Duan, Y. Li et al. 2021. Insights from Y-STRs: Forensic characteristics, genetic affinities, and linguistic classifications of Guangdong Hakka and She groups. *Front. Genet.* 12:676917.

Nei, M., and F. Tajima. 1981. DNA polymorphism detectable by restriction endonucleases. *Genetics* 97:145–163.

Reynolds, J., B. S. Weir, and C. C. Cockerham. 1983. Estimation of the coancestry coefficient: Basis for a short-term genetic distance. *Genetics* 105:767–779.

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

Song, M., Z. Wang, Y. Zhang et al. 2019. Forensic characteristics and phylogenetic analysis of both Y-STR and Y-SNP in the Li and Han ethnic groups from Hainan Island of China. *Forensic Sci. Int. Genet.* 39:e14–e20.

Su, D. 2017. The inheritance path and enlightenment of Dongshan Hakka tradition in Chengdu. *Chengdu Univ. J. Soc. Sci.* 4:36–40.

Subramanian, B., S. Gao, M. J. Lercher et al. 2019. Evolview v3: A webserver for visualization, annotation, and management of phylogenetic trees. *Nucleic Acids Res.* 47:W270–W275.

Tan, Y. 2008. *Hakka Illustrated Record*. Guangdong, CN: Lingnan Art Publishing House.

Wang, M., D. Yuan, X. Zou 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, Z. H. 1994. *History of Nationalities in China.* Beijing, CN: China Social Science Press.

Wen, B., H. Li, S. Gao et al. 2005. Genetic structure of Hmong-Mien speaking populations in East Asia as revealed by mtDNA lineages. *Mol. Biol. Evol.* 22:725–734.

Yin, C., Z. He, Y. Wang et al. 2022. Improving the regional Y-STR haplotype resolution utilizing haplogroup-determining Y-SNPs and the application of machine learning in Y-SNP haplogroup prediction in a forensic Y-STR database: A pilot study on male Chinese Yunnan Zhaoyang Han population. *Forensic Sci. Int. Genet.* 57:102659.

Yin, C., K. Su, Z. He et al. 2020. Genetic reconstruction and forensic analysis of Chinese Shandong and Yunnan Han populations by co-analyzing Y chromosomal STRs and SNPs. *Genes* 11:743.

Zhan, J. 2013. Research on the Huguang Assembly Hall buildings sited on emigration route of "Huguang to Sichuan" on Ming and Qing Dynasty. Master's thesis, Huazhong University of Science and Technology.

Zhang, Y., R. Zhang, M. Li et al. 2020. Genetic polymorphism of both 29 Y-STRs and 213 Y-SNPs in Han populations from Shandong Province, China. *Leg. Med. (Tokyo)* 47:101738.

Supplementary Table S1. Accession Numbers and Abbreviations of the Studied Han and Hakka Populations in Sichuan and Other Reference Populations

Populations	Accession number	Sum of haplotype*	Languages/Dialect
Dongshan, Sichuan, China_Hakka	YA004741 (in this study)	162	Sino-Tibetan languages/Chinese/Hakka
East Sichuan, China_Han	YA004694-2 (in this study)	22	Sino-Tibetan languages/Chinese/Sichuanese Mandarin
North Sichuan, China_Han	YA004694-2 (in this study)	21	Sino-Tibetan languages/Chinese/Sichuanese Mandarin
South Sichuan, China_Han	YA004694-2 (in this study)	135	Sino-Tibetan languages/Chinese/Sichuanese Mandarin
West Sichuan, China [Han]	YA004694	131	Sino-Tibetan languages/Chinese/Sichuanese Mandarin
Anhui, China [Han]	YA004567, YA004623, YA004648	3110	Sino-Tibetan languages/Chinese/Hui
Fujian, China [Han]	YA004711	74	Sino-Tibetan languages/Chinese/Hokkien
Guangdong, China [Han]	YA004330, YA004709	776	Sino-Tibetan languages/Chinese/Cantonese
Jiangxi, China [Han]	YA004570	1222	Sino-Tibetan languages/Chinese/Gan
Longyan, China [Han]	YA004676	875	Sino-Tibetan languages/Chinese/Hokkien
Shaanxi, China [Han]	YA004673	386	Sino-Tibetan languages/Chinese/Standard Chinese
Hunan, China [Miao]	YA004596	629	Sino-Tibetan languages/Hmong-Mien/Mien

*Sum of haplotype using for AMOVA in each population

Supplementary Table S2. The Binary Genetic Markers Used to Define the Haplogroup Branches in the Training and Testing Datasets (n = 3248)

Consolidated haplogroup branches	Detailed haplogroups	Binary genetic markers	Sum of detailed haplogroups
C2b1 (n = 101)	C2b1	Z1338	73
	C2b1a1	CTS2657	3
	C2b1b	F845	25
D1a1 (n = 110)	D1a1a	M15	7
	D1a1a1a	N1	82
	D1a1b1~	P47	21
N1a1 (n = 30)	N1a1	M46	19
	N1a1a	M178	11
N1a2 (n = 17)	N1a2	F1008	17
O1a1 (n = 24)	O1a1a	P203.1	2
	O1a1a1a	F140	1
	O1a1a1a1	F78	8
	O1a1a1b	SK1568	2
	O1a1a1b1b	Z23392	4
	O1a1a2a	CTS52	2
	O1a1a2a1a	Z23266	5
O2a1 (n = 965)	O2a1	KL1	731
	O2a1b	IMS-JST002611	66
	O2a1b1a1a1a	F11	156
	O2a1b1a1a1a1a1a1	F17	4
	O2a1b1a1a1a1a1a1a1a	F1095	2
	O2a1b1a1a1a1a1a1a2	CTS7501	1
	O2a1b1a1a1a1a1a1b1	F793	5
O2a2 (n = 1990)	O2a2	P201	100
	O2a2a	M188	61
	O2a2a1	F2588	11
	O2a2a1a1a	M159	2
	O2a2a1a2	M7	169
	O2a2b	P164	601
	O2a2b1	M134	1
	O2a2b1a1	M117	854
	O2a2b1a1a	M133	6
	O2a2b1a1a1a1	F438	5
	O2a2b1a1a1a3a	Z25853	4
	O2a2b1a1a1a4a	CTS4658	4
	O2a2b1a1a1b	CTS7634	6
	O2a2b1a2a	F444	149
	O2a2b1a2a1a	F46	10
	O2a2b1a2a1a1b	F2887	2
	O2a2b1a2a1a1b1b1	A9472	3
	O2a2b1a2a1a1b1b2b2a	CTS335	2
R1a1 (n = 11)	R1a1a1b2	Z93	1
	R1a1a1b2a	F3105	2
	R1a1a1b2a2	Z2124	6
	R1a1a1b2a2b	S4576	2

Supplementary Table S3. Haplotype Frequencies of 41-Y STR of Dongshan Hakka Population in Sichuan Province (n = 171)

HT*	Freq.	Count	HT*	Freq.	Count	HT*	Freq.	Count	HT*	Freq.	Count
H1	0.005848	1	H43	0.005848	1	H85	0.005848	1	H127	0.005848	1
H2	0.005848	1	H44	0.005848	1	H86	0.005848	1	H128	0.005848	1
H3	0.005848	1	H45	0.005848	1	H87	0.005848	1	H129	0.005848	1
H4	0.005848	1	H46	0.005848	1	H88	0.005848	1	H130	0.005848	1
H5	0.005848	1	H47	0.005848	1	H89	0.011696	2	H131	0.005848	1
H6	0.005848	1	H48	0.005848	1	H90	0.005848	1	H132	0.005848	1
H7	0.005848	1	H49	0.005848	1	H91	0.005848	1	H133	0.005848	1
H8	0.005848	1	H50	0.005848	1	H92	0.005848	1	H134	0.005848	1
H9	0.005848	1	H51	0.005848	1	H93	0.005848	1	H135	0.005848	1
H10	0.005848	1	H52	0.005848	1	H94	0.005848	1	H136	0.005848	1
H11	0.005848	1	H53	0.005848	1	H95	0.005848	1	H137	0.005848	1
H12	0.011696	2	H54	0.005848	1	H96	0.005848	1	H138	0.005848	1
H13	0.005848	1	H55	0.005848	1	H97	0.005848	1	H139	0.005848	1
H14	0.005848	1	H56	0.005848	1	H98	0.005848	1	H140	0.005848	1
H15	0.005848	1	H57	0.005848	1	H99	0.011696	2	H141	0.005848	1
H16	0.005848	1	H58	0.005848	1	H100	0.005848	1	H142	0.005848	1
H17	0.005848	1	H59	0.005848	1	H101	0.005848	1	H143	0.005848	1
H18	0.005848	1	H60	0.005848	1	H102	0.005848	1	H144	0.005848	1
H19	0.005848	1	H61	0.005848	1	H103	0.005848	1	H145	0.005848	1
H20	0.005848	1	H62	0.005848	1	H104	0.005848	1	H146	0.005848	1
H21	0.005848	1	H63	0.011696	2	H105	0.005848	1	H147	0.005848	1
H22	0.005848	1	H64	0.005848	1	H106	0.005848	1	H148	0.005848	1
H23	0.005848	1	H65	0.005848	1	H107	0.005848	1	H149	0.005848	1
H24	0.005848	1	H66	0.005848	1	H108	0.005848	1	H150	0.005848	1
H25	0.005848	1	H67	0.005848	1	H109	0.005848	1	H151	0.005848	1
H26	0.005848	1	H68	0.005848	1	H110	0.005848	1	H152	0.005848	1
H27	0.005848	1	H69	0.005848	1	H111	0.005848	1	H153	0.005848	1
H28	0.005848	1	H70	0.011696	2	H112	0.005848	1	H154	0.005848	1
H29	0.005848	1	H71	0.005848	1	H113	0.005848	1	H155	0.005848	1
H30	0.005848	1	H72	0.005848	1	H114	0.005848	1	H156	0.005848	1
H31	0.005848	1	H73	0.005848	1	H115	0.005848	1	H157	0.005848	1
H32	0.005848	1	H74	0.005848	1	H116	0.005848	1	H158	0.005848	1
H33	0.005848	1	H75	0.005848	1	H117	0.005848	1	H159	0.005848	1
H34	0.005848	1	H76	0.005848	1	H118	0.005848	1	H160	0.005848	1
H35	0.005848	1	H77	0.005848	1	H119	0.005848	1	H161	0.005848	1
H36	0.005848	1	H78	0.005848	1	H120	0.005848	1	H162	0.005848	1
H37	0.005848	1	H79	0.005848	1	H121	0.005848	1	H163	0.005848	1
H38	0.005848	1	H80	0.005848	1	H122	0.005848	1	H164	0.005848	1
H39	0.005848	1	H81	0.005848	1	H123	0.005848	1	H165	0.005848	1
H40	0.005848	1	H82	0.005848	1	H124	0.005848	1	H166	0.005848	1
H41	0.005848	1	H83	0.005848	1	H125	0.005848	1			
H42	0.005848	1	H84	0.005848	1	H126	0.005848	1			

*HT represents Haplotype. Among 171 males of Dongshan Hakka, a total of 166 haplotypes were found, of which 161 were unique. The overall HD and DC reached 0.9997 and 0.9708, respectively.

Supplementary Table S4. Haplotype Frequencies of 41-Y STR of Han Population in Sichuan Province (n = 182)

HT*	Freq.	Count	HT*	Freq.	Count	HT*	Freq.	Count	HT*	Freq.	Count
H1	0.005495	1	H45	0.005495	1	H89	0.005495	1	H133	0.005495	1
H2	0.005495	1	H46	0.005495	1	H90	0.005495	1	H134	0.005495	1
H3	0.005495	1	H47	0.005495	1	H91	0.005495	1	H135	0.005495	1
H4	0.005495	1	H48	0.005495	1	H92	0.005495	1	H136	0.005495	1
H5	0.005495	1	H49	0.005495	1	H93	0.010989	2	H137	0.005495	1
H6	0.005495	1	H50	0.005495	1	H94	0.010989	2	H138	0.005495	1
H7	0.005495	1	H51	0.005495	1	H95	0.005495	1	H139	0.005495	1
H8	0.005495	1	H52	0.005495	1	H96	0.005495	1	H140	0.005495	1
H9	0.005495	1	H53	0.005495	1	H97	0.005495	1	H141	0.005495	1
H10	0.005495	1	H54	0.005495	1	H98	0.005495	1	H142	0.005495	1
H11	0.005495	1	H55	0.005495	1	H99	0.005495	1	H143	0.005495	1
H12	0.005495	1	H56	0.005495	1	H100	0.010989	2	H144	0.005495	1
H13	0.005495	1	H57	0.005495	1	H101	0.005495	1	H145	0.005495	1
H14	0.005495	1	H58	0.005495	1	H102	0.005495	1	H146	0.005495	1
H15	0.005495	1	H59	0.005495	1	H103	0.005495	1	H147	0.005495	1
H16	0.005495	1	H60	0.005495	1	H104	0.005495	1	H148	0.005495	1
H17	0.005495	1	H61	0.005495	1	H105	0.005495	1	H149	0.005495	1
H18	0.005495	1	H62	0.010989	2	H106	0.005495	1	H150	0.005495	1
H19	0.005495	1	H63	0.005495	1	H107	0.005495	1	H151	0.005495	1
H20	0.005495	1	H64	0.005495	1	H108	0.005495	1	H152	0.005495	1
H21	0.005495	1	H65	0.005495	1	H109	0.005495	1	H153	0.005495	1
H22	0.005495	1	H66	0.005495	1	H110	0.005495	1	H154	0.005495	1
H23	0.005495	1	H67	0.005495	1	H111	0.005495	1	H155	0.005495	1
H24	0.005495	1	H68	0.005495	1	H112	0.005495	1	H156	0.005495	1
H25	0.005495	1	H69	0.005495	1	H113	0.005495	1	H157	0.005495	1
H26	0.005495	1	H70	0.005495	1	H114	0.005495	1	H158	0.005495	1
H27	0.005495	1	H71	0.005495	1	H115	0.005495	1	H159	0.005495	1
H28	0.005495	1	H72	0.005495	1	H116	0.005495	1	H160	0.005495	1
H29	0.005495	1	H73	0.005495	1	H117	0.010989	2	H161	0.005495	1
H30	0.005495	1	H74	0.005495	1	H118	0.005495	1	H162	0.005495	1
H31	0.005495	1	H75	0.005495	1	H119	0.005495	1	H163	0.005495	1
H32	0.005495	1	H76	0.005495	1	H120	0.005495	1	H164	0.005495	1
H33	0.005495	1	H77	0.005495	1	H121	0.005495	1	H165	0.005495	1
H34	0.005495	1	H78	0.005495	1	H122	0.010989	2	H166	0.005495	1
H35	0.005495	1	H79	0.005495	1	H123	0.005495	1	H167	0.005495	1
H36	0.005495	1	H80	0.005495	1	H124	0.005495	1	H168	0.005495	1
H37	0.005495	1	H81	0.005495	1	H125	0.005495	1	H169	0.005495	1
H38	0.005495	1	H82	0.005495	1	H126	0.005495	1	H170	0.005495	1
H39	0.005495	1	H83	0.005495	1	H127	0.005495	1	H171	0.005495	1
H40	0.005495	1	H84	0.005495	1	H128	0.005495	1	H172	0.005495	1
H41	0.005495	1	H85	0.005495	1	H129	0.005495	1	H173	0.005495	1
H42	0.005495	1	H86	0.005495	1	H130	0.005495	1	H174	0.005495	1
H43	0.005495	1	H87	0.005495	1	H131	0.005495	1	H175	0.005495	1
H44	0.005495	1	H88	0.005495	1	H132	0.005495	1	H176	0.005495	1

*HT represents Haplotype. Among 182 Sichuan Han males, 176 haplotypes were observed, and 170 of them

were exclusive. The overall HD and DC reached 0.9996 and 0.9670, respectively.

Supplementary Table S5. Allele Frequency and GD Values for 41 Y-STR Loci of Dongshan Hakka and Han Populations in Sichuan Province

Allele	DYS19		DYS389I		DYS389II		DYS390		DYS391		DYS392		DYS393	
	Hakka	Han	Hakka	Han	Hakka	Han	Hakka	Han	Hakka	Han	Hakka	Han	Hakka	Han
6									0.0117	0.0055				
7														
8														
9									0.0351	0.0220				
10									0.7368	0.7637	0.0234	0.0055		
11			0.0058						0.2047	0.1978	0.0702	0.1484		0.0165
12			0.5205	0.5824					0.0117	0.0110	0.0819	0.1319	0.4795	0.4780
13	0.0351	0.0495	0.3333	0.3022							0.3977	0.3187	0.3333	0.3077
13.1														
13.2														
14	0.1930	0.2418	0.1404	0.1044							0.3801	0.3791	0.1462	0.1374
14.1														
15	0.4737	0.4890		0.0110							0.0468	0.0165	0.0409	0.0604
16	0.2222	0.1593												
17	0.0702	0.0604												
18	0.0058													
18.2														
19														
19.3														
20														
20.2														
21								0.0110						
21.2														
22							0.0409	0.0659						
22.2														
23							0.3743	0.4451						
23.2														
24							0.3450	0.2802						
25							0.2164	0.1868						
26					0.0351	0.0110	0.0234	0.0110						
27					0.0936	0.0769								
28					0.2865	0.2967								
29					0.3392	0.3132								
30					0.1871	0.2033								
31					0.0409	0.0769								
32					0.0117	0.0220								
33					0.0058									
33.2														
34														
35														
36														
36.2														
36.3														
37														
37.2														
38														
38.2														
39														
40														
41														
42														
43														
44														
GD	0.6868	0.6746	0.6018	0.5615	0.7604	0.7643	0.6959	0.6877	0.4161	0.3790	0.6870	0.7190	0.6396	0.6576

Allele	DYS385		DYS438		DYS439		DYS437		DYS448		DYS456		DYS458	
	Hakka	Han	Hakka	Han	Hakka	Han	Hakka	Han	Hakka	Han	Hakka	Han	Hakka	Han
6														
7														
8				0.0110										
9				0.0110										
10	0.0087	0.0027	0.7485	0.6813	0.0292	0.0714								
11	0.0494	0.1366	0.2222	0.2692	0.3509	0.3407								
12	0.1715	0.1366	0.0234	0.0220	0.4152	0.4176					0.0058			
13	0.2762	0.2623	0.0058	0.0055	0.1579	0.1319		0.0110			0.0351	0.0165		0.0055
13.1	0.0029													
13.2														
14	0.0843	0.0710			0.0409	0.0385	0.6608	0.5824			0.1579	0.1703	0.0058	0.0110
14.1														
15	0.0291	0.0628			0.0058		0.3392	0.3956		0.0055	0.5965	0.5659	0.1228	0.1758
16	0.0378	0.0437						0.0110		0.0055	0.1170	0.1758	0.1754	0.2198
17	0.0407	0.0464							0.0117	0.0220	0.0819	0.0440	0.3567	0.1978
18	0.0901	0.0683							0.2865	0.2802	0.0058	0.0220	0.2164	0.2033
18.2									0.0058					
19	0.1017	0.0956							0.3743	0.3242		0.0055	0.0819	0.1484
19.3														
20	0.0785	0.0410							0.2281	0.2637			0.0234	0.0385
20.2														
21	0.0262	0.0246							0.0877	0.0989			0.0175	
21.2														
22	0.0029	0.0082							0.0058					
22.2														
23														
23.2														
24														
25														
26														
27														
28														
29														
30														
31														
32														
33														
33.2														
34														
35														
36														
36.2														
36.3														
37														
37.2														
38														
38.2														
39														
40														
41														
42														
43														
44														
GD	0.8579	0.8670	0.3920	0.4651	0.6810	0.6894	0.4509	0.5068	0.7221	0.7406	0.6011	0.6205	0.7770	0.8212

Allele	DYS635		YGATAH4		DYS481		DYS533		DYS549		DYS570		DYS576	
	Hakka	Han	Hakka	Han	Hakka	Han	Hakka	Han	Hakka	Han	Hakka	Han	Hakka	Han
6														
7														
8														
9														
10			0.0585	0.0385			0.1287	0.1209						0.0055
11			0.3743	0.3571			0.5789	0.5330	0.1345	0.0824				
12			0.4795	0.5330			0.2690	0.2912	0.5322	0.5604				
13			0.0760	0.0659			0.0234	0.0495	0.2749	0.2967				0.0055
13.1														
13.2														
14			0.0117	0.0055				0.0055	0.0585	0.0604		0.0055		0.0110
14.1														
15											0.0117	0.0330	0.0058	
16											0.1228	0.1648	0.0877	0.1044
17						0.0055					0.2573	0.2363	0.2222	0.1868
18	0.0058										0.2398	0.2747	0.3567	0.3022
18.2														
19	0.1053	0.1264									0.1871	0.1923	0.2047	0.2637
19.3														
20	0.2398	0.2527									0.0819	0.0714	0.0877	0.0879
20.2														
21	0.3333	0.3571			0.0175	0.0165					0.0585	0.0165	0.0234	0.0330
21.2														
22	0.1696	0.1593			0.0994	0.1703					0.0351	0.0055	0.0058	
22.2														
23	0.0819	0.0769			0.3275	0.2802					0.0058		0.0058	
23.2														
24	0.0526	0.0220			0.2222	0.1868								
25	0.0117	0.0055			0.1170	0.1429								
26					0.1111	0.0879								
27					0.0526	0.0989								
28					0.0175	0.0055								
29					0.0117	0.0055								
30					0.0234									
31														
32														
33														
33.2														
34														
35														
36														
36.2														
36.3														
37														
37.2														
38														
38.2														
39														
40														
41														
42														
43														
44														
GD	0.7865	0.7650	0.6243	0.5858	0.8081	0.8238	0.5787	0.6174	0.6234	0.5907	0.8195	0.8024	0.7699	0.7887

Allele	DYS643		DYF387S1		DYS449		DYS460		DYS518		DYS627		DYS388	
	Hakka	Han	Hakka	Han	Hakka	Han	Hakka	Han	Hakka	Han	Hakka	Han	Hakka	Han
6														
7		0.0055												
8	0.0117	0.0604												
9	0.0877	0.0385					0.3099	0.3571						0.0055
10	0.1930	0.2637					0.4152	0.3681					0.1637	0.1648
11	0.5029	0.4066					0.2573	0.2363					0.0058	0.0110
12	0.1871	0.1923					0.0175	0.0385					0.7368	0.6209
13	0.0175	0.0330											0.0760	0.1538
13.1														
13.2														
14													0.0175	0.0440
14.1														
15											0.0058			
16											0.0058			
17											0.0058	0.0275		
18											0.0468	0.0769		
18.2														
19											0.0936	0.1429		
19.3														
20											0.1813	0.1758		
20.2														
21											0.2573	0.2308		
21.2														
22											0.2573	0.1978		
22.2												0.0055		
23											0.1111	0.1209		
23.2														
24											0.0175	0.0220		
25						0.0110					0.0175			
26					0.0351	0.0495								
27					0.0468	0.0440								
28					0.0760	0.0714								
29					0.0877	0.1209								
30					0.1637	0.1209								
31					0.1228	0.1758				0.0055				
32					0.2047	0.1868				0.0110				
33			0.0058	0.0055	0.1637	0.0824			0.0058	0.0165				
33.2														
34			0.0116	0.0137	0.0585	0.0604			0.0468	0.0330				
35			0.1105	0.1233	0.0234	0.0440			0.0877	0.0714				
36			0.1657	0.2000		0.0220			0.1462	0.1703				
36.2										0.0055				
36.3			0.0029											
37			0.1977	0.1863	0.0175	0.0110			0.1637	0.1209				
37.2									0.0175					
38			0.1948	0.2110					0.1579	0.2582				
38.2														
39			0.1715	0.1534					0.1696	0.1209				
40			0.0959	0.0877					0.1053	0.0604				
41			0.0378	0.0164					0.0643	0.0659				
42			0.0029	0.0027					0.0351	0.0330				
43			0.0029							0.0275				
44														
GD	0.6706	0.7259	0.8455	0.8362	0.8733	0.8873	0.6689	0.6834	0.8766	0.8633	0.8155	0.8391	0.4266	0.5647

Allele	DYS447		DYS444		DYS645		DYS557		DYS522		DYS596		DYS593	
	Hakka	Han	Hakka	Han	Hakka	Han	Hakka	Han	Hakka	Han	Hakka	Han	Hakka	Han
6														
7														
8					0.9408	0.9669								
9					0.0592	0.0331				0.0056				
10				0.0110					0.1302	0.1056				
11			0.1941	0.2155					0.3609	0.3722				
12			0.3882	0.3094					0.4024	0.4222				
13			0.2471	0.2762			0.0409	0.0330	0.1065	0.0944		0.0055		
13.1														
13.2														
14			0.1353	0.1602			0.3977	0.3791			0.3041	0.3407		
14.1														
15			0.0353	0.0276			0.2222	0.2363			0.6199	0.5824	0.3882	0.3681
16							0.1988	0.1758			0.0760	0.0659	0.4294	0.4451
17							0.0819	0.0604				0.0055	0.1647	0.1703
18		0.0055					0.0409	0.0769					0.0176	0.0165
18.2														
19							0.0175	0.0385						
19.3														
20														
20.2														
21	0.0409	0.0110												
21.2														
22	0.0175	0.0220												
22.2														
23	0.1754	0.2088												
23.2														
24	0.2632	0.2143												
25	0.1988	0.2912												
26	0.1170	0.1209												
27	0.1170	0.0934												
28	0.0468	0.0330												
29	0.0234													
30														
31														
32														
33														
33.2														
34														
35														
36														
36.2														
36.3														
37														
37.2														
38														
38.2														
39														
40														
41														
42														
43														
44														
GD	0.8332	0.8051	0.7353	0.7592	0.1120	0.0645	0.7470	0.7616	0.6836	0.6668	0.5205	0.5433	0.6412	0.6406

Allele	DYS527		DYF404S1	
	Hakka	Han	Hakka	Han
6				
7				
8				
9				
10				0.0027
11				0.0027
12			0.0667	0.1014
13			0.2174	0.2493
13.1				
13.2			0.0058	
14			0.2696	0.2822
14.1				
15			0.3072	0.2466
16			0.0986	0.0986
17	0.0029	0.0027	0.0290	0.0137
18			0.0029	0.0027
18.2				
19	0.0496	0.0522	0.0029	
19.3				
20	0.1370	0.1593		
20.2				
21	0.2099	0.1841		
21.2				
22	0.2828	0.2418		
22.2				
23	0.1808	0.2005		
23.2				
24	0.1166	0.1099		
25	0.0175	0.0357		
26	0.0029	0.0082		
27		0.0027		
28				
29		0.0027		
30				
31				
32				
33				
33.2				
34				
35				
36				
36.2				
36.3				
37				
37.2				
38				
38.2				
39				
40				
41				
42				
GD	0.8105	0.8282	0.7729	0.7793

GD: genetic diversity, GD= n (1-$\sum pi^2$) / (n-1), where n is the total number of observed allele and *pi* represents the frequency of the *i*-th allele.

Supplementary Table S6. Allelic Combination Distributions of Four Multi-copy Markers of Hakka and Han Populations in Sichuan Province

DYS385				DYF387S1			
Hakka		Han		Hakka		Han	
Haplotype	Frequency	Haplotype	Frequency	Haplotype	Frequency	Haplotype	Frequency
10, 18	0.0058	10, 17	0.0055	35, 39	0.0877	33, 33	0.0055
10, 20	0.0117	11, 11	0.0824	36, 39	0.0819	34, 38	0.0055
11, 11	0.0058	11, 12	0.0495	37, 37	0.0702	34, 39	0.0165
11, 12	0.0292	11, 13	0.0165	38, 38	0.0468	34, 40	0.0055
11, 17	0.0175	11, 16	0.0055	36, 40	0.0468	35, 35	0.0110
11, 18	0.0117	11, 18	0.0165	35, 40	0.0292	35, 36	0.0165
11, 19	0.0175	11, 19	0.0055	35, 38	0.0409	35, 37	0.0440
11, 20	0.0117	11, 20	0.0055	36, 37	0.0585	35, 38	0.1099
12, 12	0.0292	11, 21	0.0110	35, 37	0.0292	35, 38, 40	0.0055
12, 13, 20, 21	0.0058	12, 12	0.0165	36, 38	0.0936	35, 39	0.0330
12, 13.1	0.0058	12, 14	0.0055	35, 36, 37	0.0058	35, 40	0.0165
12, 14	0.0058	12, 16	0.0604	39, 39	0.0409	36, 36	0.0549
12, 15	0.0117	12, 17	0.0275	39, 40	0.0117	36, 37	0.0385
12, 16	0.0292	12, 18	0.0385	35, 41	0.0117	36, 38	0.0879
12, 17	0.0351	12, 19	0.0220	37, 41	0.0409	36, 39	0.0769
12, 18	0.0175	12, 20	0.0330	37, 40	0.0409	36, 40	0.0440
12, 19	0.0702	12, 22	0.0055	38, 40	0.0468	36, 41	0.0275
12, 20	0.0585	13, 13	0.1044	37, 38	0.0351	37, 37	0.0604
12, 21	0.0175	13, 14	0.0549	36.3, 38	0.0058	37, 38	0.0330
13, 13	0.1053	13, 14, 19, 20	0.0055	38, 39	0.0468	37, 39	0.0934
13, 14	0.0585	13, 16	0.0055	37, 39	0.0351	37, 40	0.0330
13, 15	0.0058	13, 17	0.0385	34, 34	0.0058	37, 41	0.0055
13, 16	0.0058	13, 18	0.0330	35, 35	0.0058	37, 42	0.0055
13, 17	0.0175	13, 19	0.1154	36, 36	0.0175	38, 38	0.0495
13, 18	0.1053	13, 20	0.0385	33, 40	0.0117	38, 39	0.0330
13, 19	0.0643	13, 21	0.0110	38, 41	0.0117	38, 40	0.0495
13, 20	0.0526	14, 14	0.0055	36, 41	0.0117	39, 39	0.0220
13, 21	0.0292	14, 15	0.0110	34, 38	0.0117	39, 40	0.0110
14, 14	0.0117	14, 18	0.0220	35, 43	0.0058	40, 40	0.0055
14, 16	0.0058	14, 19	0.0220	37, 38, 40	0.0058		
14, 17	0.0058	14, 21	0.0055	37, 42	0.0058		
14, 18	0.0409	14, 22	0.0055				
14, 19	0.0234	15, 15	0.0220				
14, 20	0.0058	15, 16	0.0110				
15, 16	0.0117	15, 17	0.0220				
15, 19	0.0175	15, 18	0.0055				
15, 20	0.0058	15, 19	0.0110				
15, 22	0.0058	15, 21	0.0165				
16, 16	0.0058	15, 22	0.0055				
16, 19	0.0058	16, 21	0.0055				
16, 20	0.0058	18, 18	0.0110				
17, 19	0.0058	19, 19	0.0055				

DYS385				DYF387S1			
Hakka		Han		Hakka		Han	
Haplotype	Frequency	Haplotype	Frequency	Haplotype	Frequency	Haplotype	Frequency
20, 23	0.0760	18, 23	0.0055	13, 13	0.0819	10, 13	0.0055
21, 21	0.0351	19, 19	0.0055	14, 15	0.1871	11, 14	0.0055
20, 20	0.0175	19, 20	0.0440	15, 15	0.1287	12, 12	0.0275
22, 22	0.1053	19, 21	0.0110	16, 16	0.0292	12, 13	0.0604
20, 24	0.0760	19, 23	0.0055	12, 13	0.0526	12, 14	0.0385
24, 24	0.0117	19, 24	0.0220	14, 14	0.0819	12, 15	0.0330
21, 24	0.0760	19, 25	0.0110	13, 14	0.0936	12, 16	0.0165
20, 25	0.0175	20, 20	0.0165	13, 15	0.0819	13, 13	0.0989
19, 22	0.0234	20, 21	0.0275	13, 17	0.0117	13, 14	0.1209
21, 23	0.0877	20, 22	0.0495	15, 16, 17	0.0058	13, 14, 17	0.0055
19, 19	0.0175	20, 23	0.0549	14, 14, 16, 17	0.0058	13, 15	0.0824
21, 22	0.1462	20, 24	0.0714	15, 17	0.0117	13, 16	0.0220
23, 24	0.0234	20, 25	0.0385	15, 16	0.0409	13, 18	0.0055
22, 26	0.0058	21, 21	0.0440	17, 17	0.0058	14, 14	0.0934
20, 22	0.0468	21, 22	0.0879	13, 16	0.0234	14, 15	0.1538
22, 23	0.0936	21, 23	0.0989	14, 16	0.0468	14, 16	0.0440
23, 23	0.0351	21, 24	0.0385	12, 14	0.0234	14, 17	0.0110
17, 23	0.0058	21, 25	0.0110	13.2, 14	0.0117	15, 15	0.0824
22, 24	0.0351	21, 26	0.0055	12, 15	0.0351	15, 16	0.0604
21, 25	0.0175	22, 22	0.1099	12, 16	0.0175	16, 16	0.0275
19, 21	0.0175	22, 23	0.0714	13, 18	0.0058	17, 17	0.0055
21, 22, 23	0.0058	22, 24	0.0385	14, 17	0.0058		
19, 20	0.0234	22, 25	0.0055	12, 17	0.0058		
		22, 26	0.0055	13, 19	0.0058		
		22, 27	0.0055				
		23, 23	0.0604				
		23, 24	0.0275				
		23, 25	0.0055				
		23, 26	0.0055				
		23, 29	0.0055				
		24, 24	0.0110				

Supplementary Table S7. Copy Number Variants Observed in Hakka and Han Populations in Sichuan Province

Genotype	Locus	Count	Sample Origin
12, 13, 20, 21	DYS385	1	Dongshan Hakka
13, 14, 19, 20	DYS385	1	Sichuan Han
35, 36, 37	DYF387S1	1	Dongshan Hakka
35, 38, 40	DYF387S1	1	Sichuan Han
37, 38, 40	DYF387S1	1	Dongshan Hakka
21, 22, 23	DYS527	1	Dongshan Hakka
13, 14, 17	DYF404S1	1	Sichuan Han
14, 14, 16, 17	DYF404S1	1	Dongshan Hakka
15, 16, 17	DYF404S1	1	Dongshan Hakka

Supplementary Table S8. Pairwise *R*ST value Estimates (below the diagonal) and Corresponding *p* Value (above the diagonal) among the Dongshan Hakka Population and 11 Other Populations

Population	West Sichuan, China [Han]	**East Sichuan, China _Han**	**North Sichuan, China _Han**	**South Sichuan, China _Han**	**Sichuan, China _Hakka**	Anhui, China [Han]	Fujian, China [Han]	Guangdong, China [Han]	Jiangxi, China [Han]	Longyan, China [Han]	Shaanxi, China [Han]	Hunan, China [Miao]
West Sichuan, China [Han]	-	0.3829	0.5504	0.2909	0.0054	0.0006	0.0016	0.0236	0.0000	0.0000	0.0000	0.0578
East Sichuan, China_Han	0.0013	-	0.7044	0.6735	0.4346	0.5959	0.4372	0.6362	0.4834	0.6502	0.4617	0.5680
North Sichuan, China_Han	-0.0036	-0.0131	-	0.8890	0.0987	0.5775	0.2980	0.8316	0.4361	0.4988	0.8138	0.7930
South Sichuan, China_Han	0.0012	-0.0070	-0.0129	-	0.0105	0.1332	0.0816	0.8707	0.0560	0.1242	0.0516	0.0900
Dongshan, Sichuan, China_Hakka	0.0119	-0.0007	0.0152	0.0099	-	0.0006	0.0077	0.0003	0.0034	0.0040	0.0000	0.0000
Anhui, China [Han]	0.0095	-0.0040	-0.0039	0.0017	0.0076	-	0.0509	0.0000	0.0000	0.0000	0.0000	0.0000
Fujian, China [Han]	0.0239	-0.0014	0.0044	0.0071	0.0159	0.0060	-	0.0086	0.1934	0.2099	0.0010	0.0002
Guangdong, China [Han]	0.0046	-0.0050	-0.0093	-0.0018	0.0096	0.0026	0.0107	-	0.0000	0.0000	0.0002	0.0001
Jiangxi, China [Han]	0.0139	-0.0018	-0.0009	0.0034	0.0073	0.0017	0.0023	0.0042	-	0.0019	0.0000	0.0000
Longyan, China [Han]	0.0160	-0.0056	-0.0023	0.0022	0.0069	0.0031	0.0021	0.0054	0.0022	-	0.0000	0.0000
Shaanxi, China [Han]	0.0176	-0.0011	-0.0089	0.0040	0.0220	0.0084	0.0186	0.0056	0.0112	0.0105	-	0.0000
Hunan, China [Miao]	0.0040	-0.0041	-0.0097	0.0031	0.0176	0.0101	0.0267	0.0048	0.0155	0.0164	0.0096	-

Pairwise genetic distances (*R*ST value estimates, below the diagonal) were obtained by Analysis of MOlecular VAriance (AMOVA). Significance of genetic distances (above the diagonal) was tested by 10,000 permutations. Significant differentiation test *p*-values were marked in red ($p = 0.05/66=0.000758$ after Bonferroni correction).

Supplementary Table S9. Confusion Matrix and Statistics for Prediction on the Testing Set of Y-DNA Database from Chinese Han Populations

Prediction[#]	Truth[*]							
	C2b1	D1a1	N1a1	N1a2	O1a1	O2a1	O2a2	R1a1
C2b1	100	0	0	0	0	1	3	0
D1a1	0	108	0	0	0	0	1	0
N1a1	0	0	28	0	0	0	0	0
N1a2	0	0	2	17	0	0	1	0
O1a1	0	0	0	0	22	1	3	0
O2a1	0	0	0	0	2	937	37	0
O2a2	0	1	0	0	0	17	1926	0
R1a1	0	0	0	0	0	0	0	11

[#]The vertical axis represents the predicted label (haplogroup) while *the abscissa axis represents the true label (haplogroup).

Supplementary Table S10. Overall Statistics for Prediction on the Testing Set of Y-DNA Database from Chinese Han Populations

Haplogroups	C2b1	D1a1	N1a1	N1a2	O1a1	O2a1	O2a2	R1a1
Sensitivity	1.0000	0.9908	0.9333	1.0000	0.9167	0.9801	0.9772	1.0000
Specificity	0.9987	0.9997	1.0000	0.9991	0.9987	0.9828	0.9856	1.0000
Pos Pred Value	0.9615	0.9908	1.0000	0.8500	0.8462	0.9600	0.9907	1.0000
Neg Pred Value	1.0000	0.9997	0.9994	1.0000	0.9994	0.9915	0.9647	1.0000
Prevalence	0.0311	0.0339	0.0093	0.0053	0.0075	0.2971	0.6125	0.0034
Detection Rate	0.0311	0.0336	0.0087	0.0053	0.0068	0.2912	0.5985	0.0034
Detection Prevalence	0.0323	0.0339	0.0087	0.0062	0.0081	0.3033	0.6041	0.0034
Balanced Accuracy	0.9994	0.9953	0.9667	0.9995	0.9577	0.9814	0.9814	1.0000

Figure Captions

Figure 1. MDS analysis of Dongshan Hakka and 11 other populations based on R_{ST} values.

Figure 2. Population substructure reconstruction for the male of Hakka and Han populations in Sichuan Province (n = 338) based on LDA method. All samples with null alleles were excluded for analysis.

Figure 3. Phylogenetic relationships among Dongshan Hakka and other reference populations based on the pairwise R_{ST} genetic distances.

Figure 4. Y-chromosome haplogroups and their frequency distribution of the four studied populations in Sichuan Province. OpenStreetMap (© OpenStreetMap contributors; see openstreetmap.org/copyright for license information) was used to create this map.

Figure 5. Median-joining networks of Y-haplogroup O2a1 and O2a2 constructed for the four studied populations based on 33 Y-STR loci. The lines between the circles are proportional to the number of mutational steps, with the area of the circles proportional to the haplotype frequency.

Supplementary Figure S1. Geographic distribution of sampling sites. Male samples (n = 353) were collected from Sichuan, Southwest China. The sampling sites of Han populations are separately located in three different directions (South, East, North) inside the Sichuan Basin. The pin of Magenta, Cyan, and Blue color indicate the three Han populations, respectively. Orange pin presents the Dongshan Hakka population. Green pins indicate the Sichuan Han in the previous study (YA004694). OpenStreetMap (© OpenStreetMap contributors; see openstreetmap.org/copyright for license information) was used to create this map.

Supplementary Figure S2. The genetic diversities of the 41 Y-STR loci in the populations of Dongshan Hakka (n = 171) and Sichuan Han (n = 182) genotyped using SureID® PathFinder Plus. Broken lines represent genetic diversity of 0.50 and 0.70.

Figure 1.

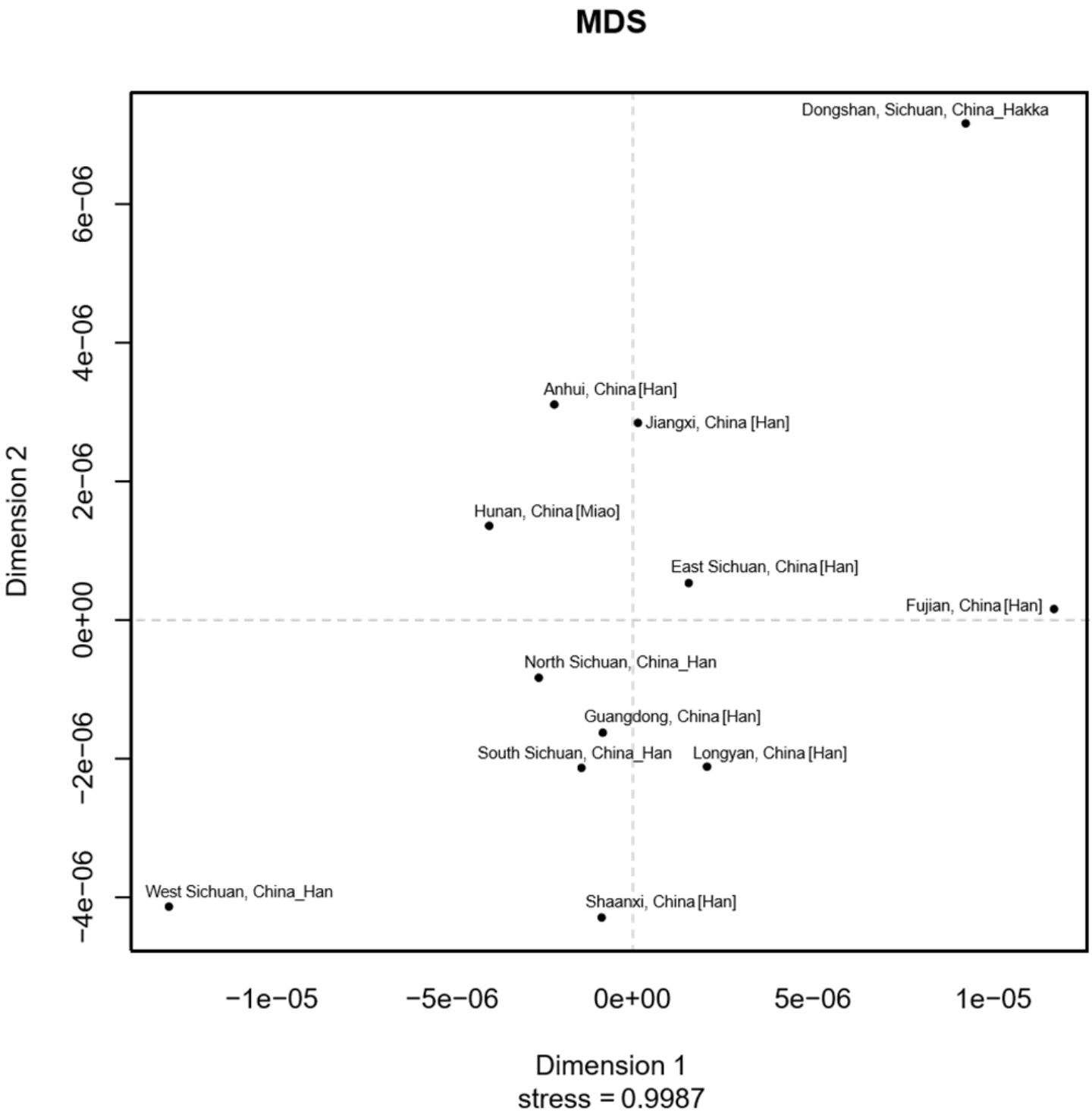


Figure 2.

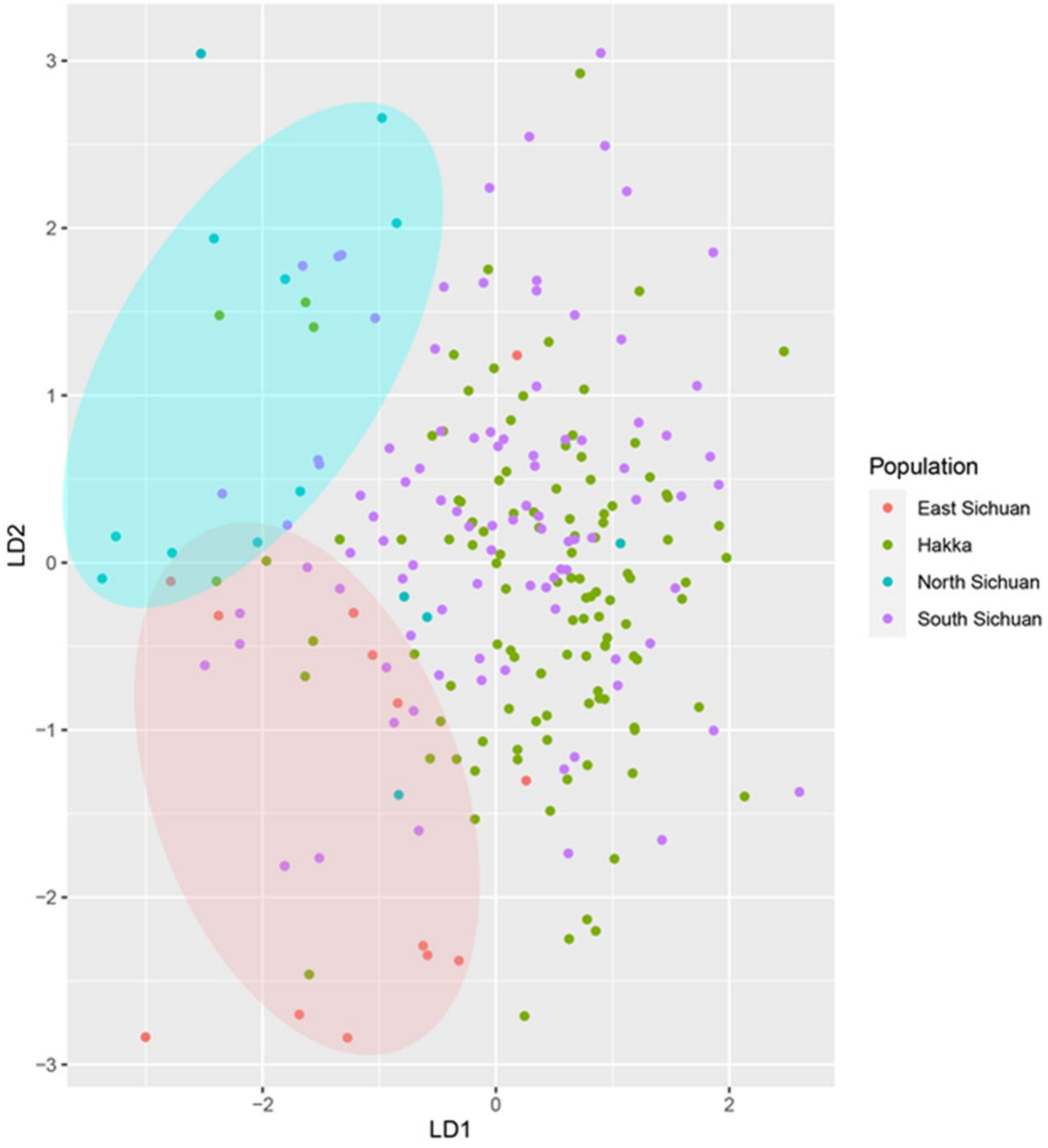


Figure 3.

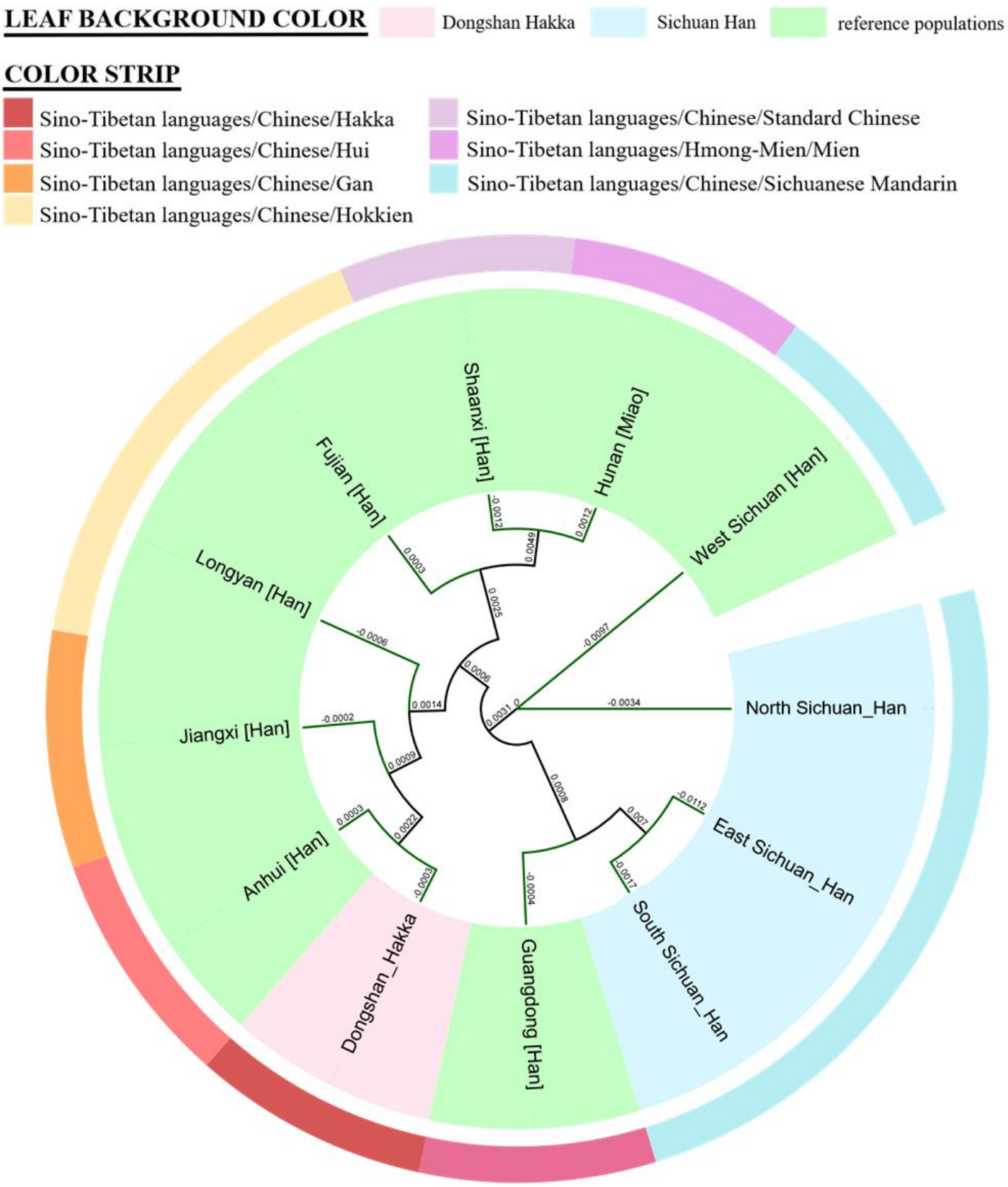

LEAF BACKGROUND COLOR
Dongshan Hakka
Sichuan Han
reference populations
COLOR STRIP
Sino-Tibetan languages/Chinese/Hakka
Sino-Tibetan languages/Chinese/Hui
Sino-Tibetan languages/Chinese/Gan
Sino-Tibetan languages/Chinese/Hokkien
Sino-Tibetan languages/Chinese/Standard Chinese
Sino-Tibetan languages/Hmong-Mien/Mien
Sino-Tibetan languages/Chinese/Sichuanese Mandarin
Shaanxi [Han]
Hunan [Miao]
West Sichuan [Han]
Fujian [Han]
Longyan [Han]
Jiangxi [Han]
Anhui [Han]
Dongshan_Hakka
Guangdong [Han]
South Sichuan_Han
East Sichuan_Han
North Sichuan_Han
-0.0012
0.0012
0.0004
0.0003
0.0025
-0.0006
0.0006
0.0014
-0.0097
-0.0034
0.0031
-0.0002
0.0009
0.0003
0.0022
-0.0003
0.0008
-0.0004
0.007
-0.0017
-0.0112

Figure 4.

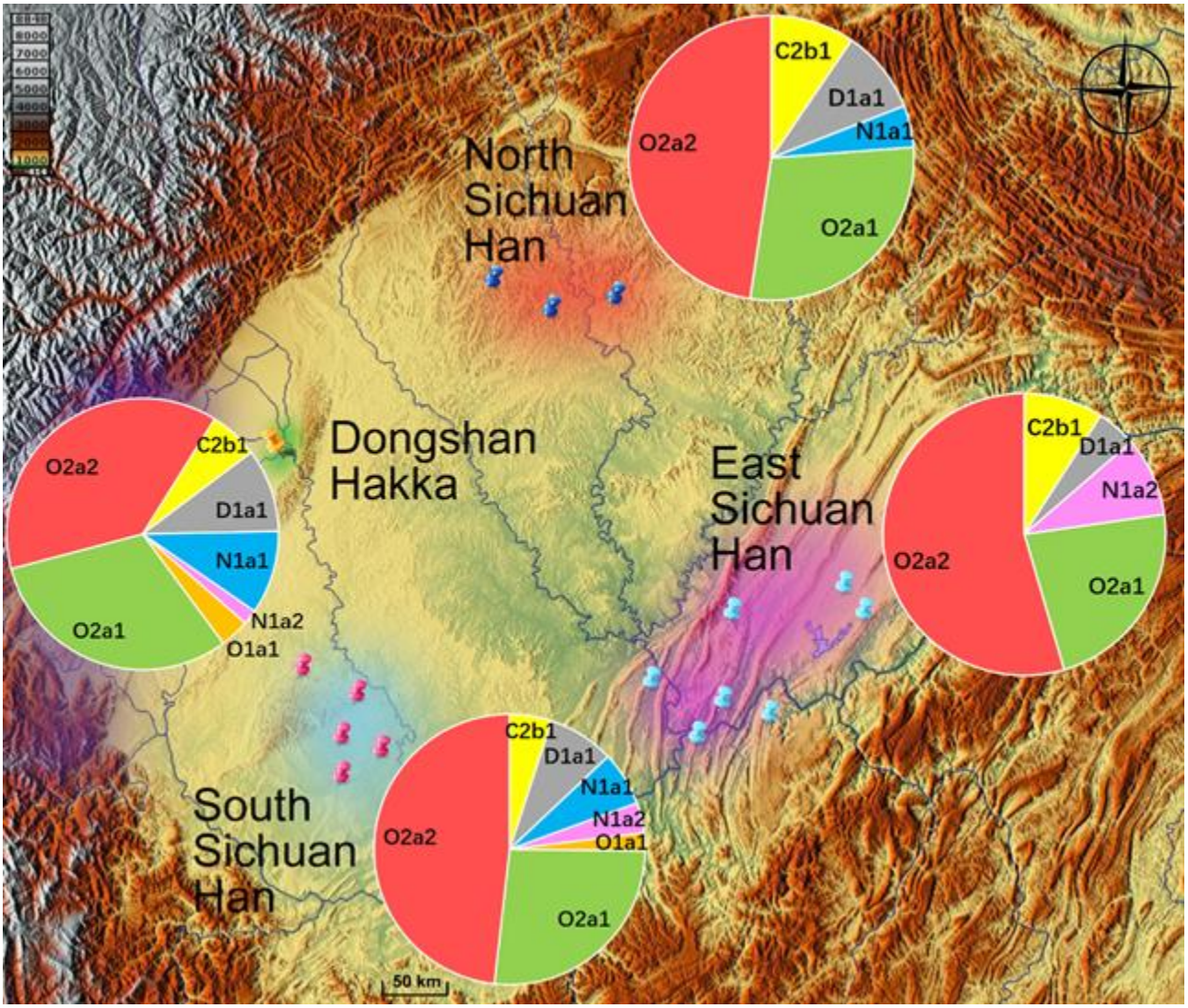
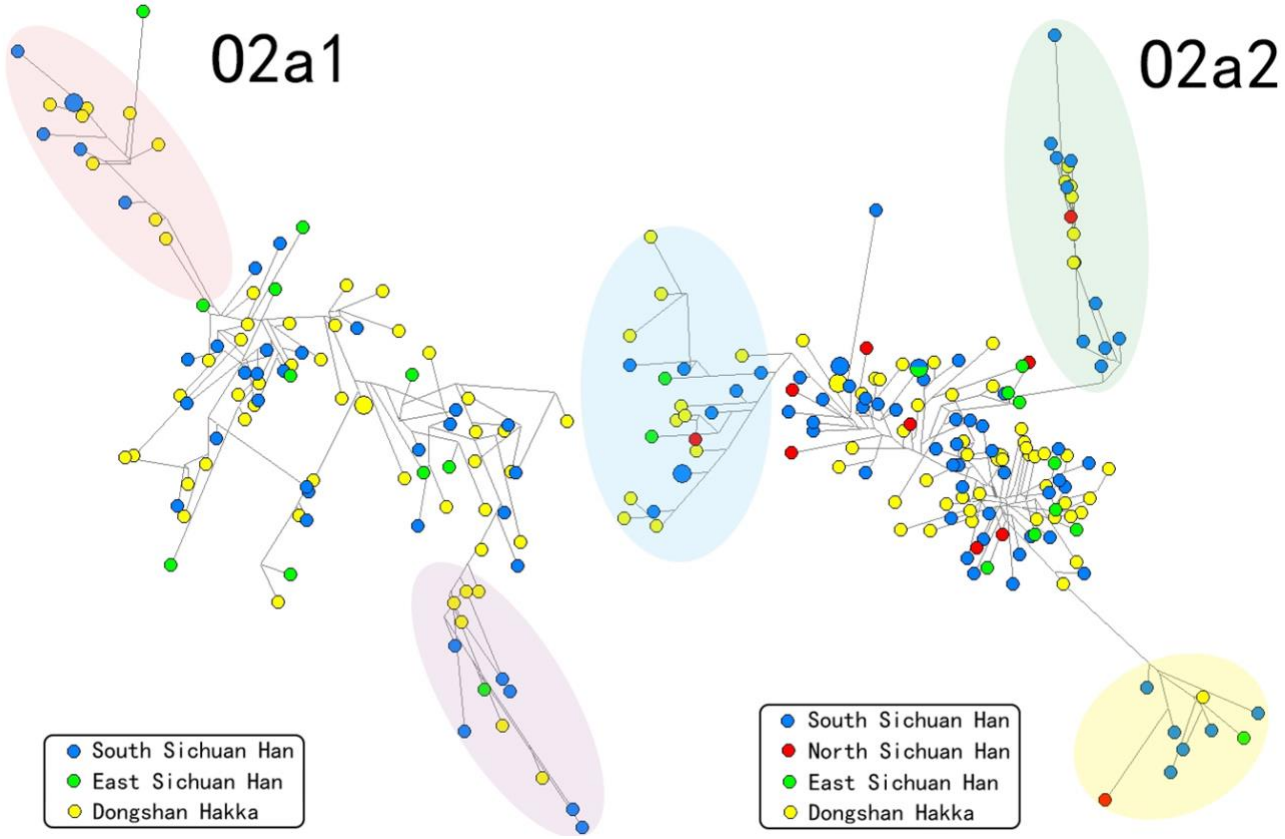
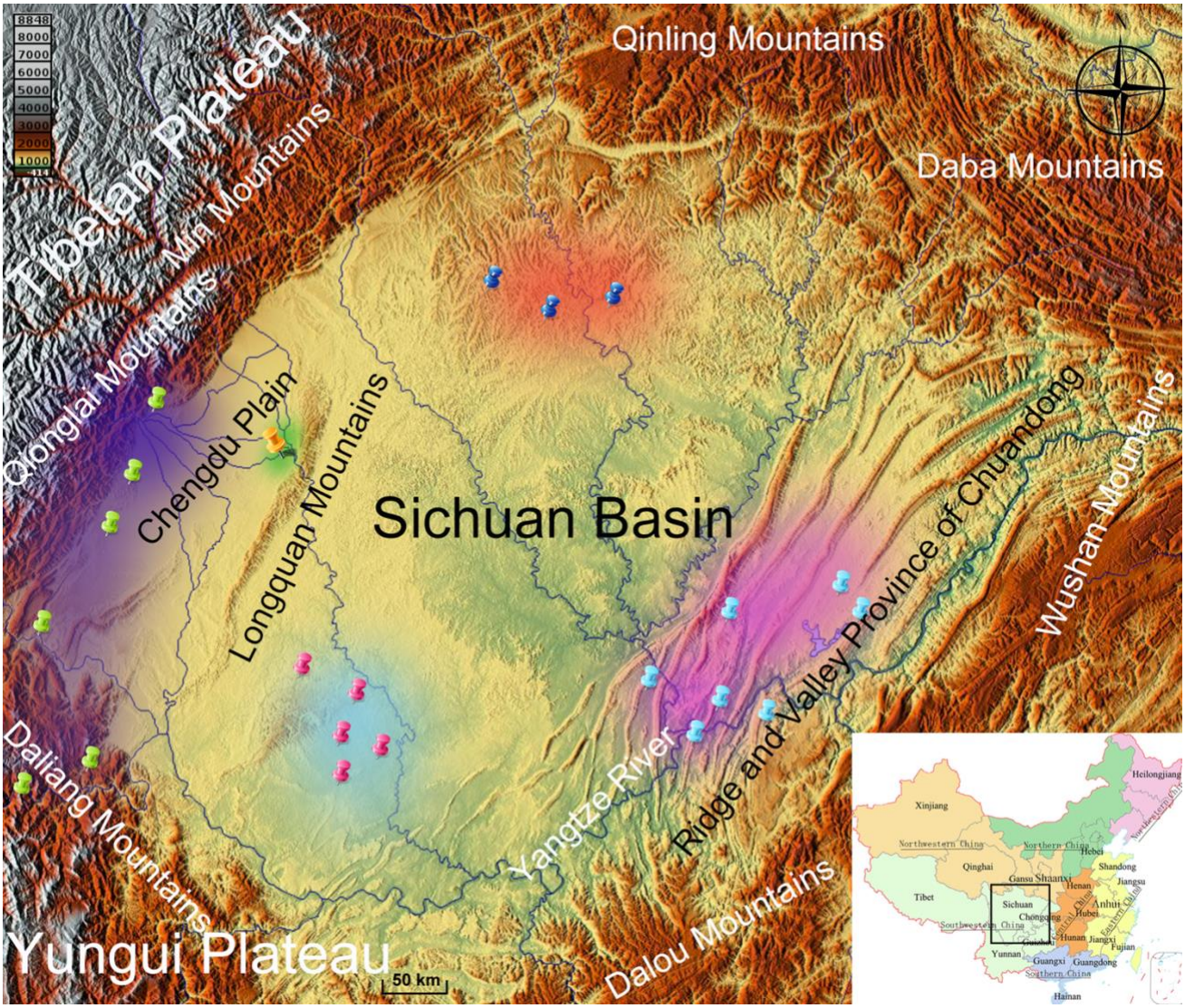

North Sichuan Han
C2b1
D1a1
N1a1
O2a2
O2a1
Dongshan Hakka
C2b1
D1a1
N1a1
N1a2
O1a1
O2a2
O2a1
East Sichuan Han
C2b1
D1a1
N1a2
O2a2
O2a1
South Sichuan Han
C2b1
D1a1
N1a1
N1a2
O1a1
O2a2
O2a1
50 km

Figure 5.



Supplementary Figure S1.



Supplementary Figure S2.