Genetic Variability and Phylogenetic Analysis of Han Population from Guanzhong Region of China based on 21 non-CODIS STR Loci

Yu-Dang Zhang1,2*, Xiao-Li Tang3*, Hao-Tian Meng1,2, Hong-Dan Wang4, Rui Jin2,5, Chun-Hua Yang2,6, Jiang-Wei Yan7, Guang Yang8, Wen-Juan Liu2, Chun-Mei Shen9 & Bo-Feng Zhu1

1Research Center of Stomatology, Stomatological Hospital, Xi’an Jiaotong University, Xi’an, 710004, P. R. China, 2Xi’an Jiaotong University Health Science Center, Xi’an 710061, P. R. China, 3Department of Biochemistry, Medical college of Nanchang University, Nanchang, 330001, P. R. China, 4Medical Genetic Institute of Henan Province, Henan Provincial People’s Hospital, People’s Hospital of Zhengzhou University, Zhengzhou 450003, P. R. China, 5Department of Radiology, The Second Affiliated Hospital of Xi’an Jiaotong University, Xi’an, 710004, P. R. China, 6People’s hospital of Arong Banner, Hulun Buir City, 162750, P. R. China, 7Key Laboratory of Genome Sciences, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing 100101, P. R. China, 8Department of Pathology and Laboratory Medicine Loma Linda University Medical Center, California, 92354, USA, 9College of Life Sciences, Shaanxi Normal University, Xi’an, 710062, P. R. China.

In the present study, we presented the population genetic data and their forensic parameters of 21 non-CODIS autosomal STR loci in Chinese Guanzhong Han population. A total of 166 alleles were observed with corresponding allelic frequencies ranging from 0.0018 to 0.5564. No STR locus was observed to deviate from the Hardy-Weinberg equilibrium and linkage disequilibriums after applying Bonferroni correction. The cumulative power of discrimination and probability of exclusion of all the 21 STR loci were 0.99999999999999999993814 and 0.999998184, respectively. The results of genetic distances, phylogenetic trees and principal component analysis revealed that the Guanzhong Han population had a closer relationship with Ningxia Han, Tujia and Bai groups than other populations tested. In summary, these 21 STR loci showed a high level of genetic polymorphisms for the Guanzhong Han population and could be used for forensic applications and the studies of population genetics.

China is an ancient country with 5,000-year-long civilization and has the largest population in the world, about 1.371 billion in the sixth national population census of China in 2010. As the biggest one of the 56 ethnic groups and with a population of approximately 1.226 billion, the Han population is widespread across China. Their spoken and written language is Chinese, one branch of the Sino-Tibetan language family. Chu et al. constructed the phylogenies using the neighbor-joining method based on difference population data for short tandem repeat (STR) loci and concluded that there was the distinction between southern and northern populations in China1. For Chinese Han population, previous population genetic studies based on STRs or single nucleotide polymorphisms (SNPs) have shown that the Chinese Han population was intricately sub-structured and clustered roughly to two (northern Han and southern Han)2-3 or three (northern Han, central Han and southern Han) subgroups4. So, it is of significance to further clarify the genetic structure of Chinese Han populations from different regions.

Guanzhong region, literally means “within the passes” in Chinese, is located in the middle of the Chinese mainland and includes the cities of Xi’an, Tongchuan, Baoji, Xianyang and Weinan in Shaanxi province, China. There are several ethnic groups, mainly including Han, Hui and Manchu nationalities living together in the region. Shen et al. reported that the Guanzhong Han population had the close genetic relationship with the northern and southern Han populations using genetic distance measurements, neighbor-joining dendrograms and principal component analysis (PCA) base on different HLA loci5.

STRs have been the most widely used in forensic science and population genetics. In order to provide more genetic information and increase the power of discrimination (PD) and probability of exclusion (PE), more novel STR loci with high genetic polymorphisms were integrated into one fluorescence-labeled multiplex amplification
system. And, it is necessary to analyze the allelic distribution of STR loci before used in forensic applications. We have so far reported population data for a panel of 21 STR loci, and these STR loci demonstrated tremendous potential for forensic applications. In the present study, we first aimed to present the population genetic data and forensic parameters of the Chinese Guanzhong Han (Northern Han in geography) with a panel of 21 non-CODIS autosomal STRs. Moreover, we investigated the genetic relationships and population differentiations between Guanzhong Han and other Chinese groups.

Methods

Populations and DNA extraction. Blood samples were randomly collected from 275 unrelated individual of the Han Chinese living in Guanzhong region, Shaanxi province, China. Before getting involved in the study, all the participants signed the written informed consents for the sample collections and succedent analyses. This study was conducted according to the humane and ethical research principles and approved by the ethical committee of Xi’an Jiaotong University Health Science Center, China. The genomic DNA was extracted from blood-stained samples using the Chelex-100 method as described by Walsh et al.

Genotyping results of the 21 STR loci from 10 Chinese groups were chosen for population comparison, including Mongolian (n = 86) from Inner Mongolia autonomous region, Bai (n = 106) from Yunnan province, Kazak (n = 114) from Xinjiang autonomous region, Ningxia Han (Northern Han) (n = 202) from Ningxia autonomous region, Russian (n = 114) from Inner Mongolia autonomous region, Tibetan (n = 104) from Tibet autonomous region, Tuja (n = 107) from Hubei province, Uigur (n = 218) from Xinjiang autonomous region, Yi (n = 110) from Yunnan province, Salar (n = 120) from Qinghai province. The geographical locations of the reference populations were shown in Figure 1.

PCR amplification and STR typing. A panel of STRs were amplified in a single reaction using the AGCU 21+1 STR system (AGCU ScienTech Incorporation, Wuxi, Jiangsu, China), according to the manufacturer’s instructions. The PCR products were separated and detected by capillary electrophoresis on the ABI 3130xl Genetic Analyzer (Applied Biosystems, Foster City, CA, USA). The STR typing results were obtained by comparing to the 21+1 Allelic Ladder using the program GeneMapper® ID-X v1.3 (Applied Biosystems, Foster City, CA, USA). Control DNA from 9947A cell line (Promega Corporation, Madison, WI, USA) was typed for quality control. All laboratory procedures were in accordance with the laboratory internal control standards.

Statistical analyses. Allelic frequencies and forensic parameters were calculated using the modified Powerstats v1.217. The Genepop v4.0.10 (http://genepop.curtin.edu.au/) was utilized to estimate the linkage disequilibriums (LDs) for all pair-wise STR loci. To estimate the inter-population differentiations between the Guanzhong Han and 10 reference populations in China, the locus-by-locus Fst, associated p and overall Fst values were calculated using the method of analysis of molecular variance (AMOVA) by the software ARLEQUIN v3.1 (http://cmpg.unibe.ch/software/arlequin3) and the D values were calculated using the DISPAN program. To visually estimate the genetic relationships between the Guanzhong Han and reference populations, we performed two kinds of phylogenetic trees using the software MEGA v5 with the unweighted pair-group method with arithmetic means (UPGMA) based on Ds distances and the software PHYLIP v3.6 by a bootstrap-over-loci method with 1,000 replicates based on allelic frequencies, respectively. A PCA plot was conducted with MATLAB 2007a (MathWorks Inc., USA) based on allelic frequencies of 21 STRs. The existence of significant LD among STRs has an impact on some subsequent analyses, including Ds calculation and MEGA, so the STR loci which observed to be in significant LD with one or more other loci would be removed in the analyses mentioned above.

Results and Discussion

The typing results of the 21 STR loci from the Guanzhong Han population were listed in supplemental Table 1, and the allelic frequencies and forensic parameters were shown in Table 1. A total of

Figure 1 | The geographical locations of the Guanzhong Han and 10 reference groups in China. The map was created in matlab R2013b software (MathWorks Inc., USA).
Table 1 | The allelic frequencies and statistical parameters for the 21 STR loci in Han population from Guanzhong region, Shaanxi, China (n = 275)

| Allele | D6S474 | D12ATA63 | D22S1045 | D10S1248 | D1S1677 | D11S4463 | D1S1627 | D3S4529 | D2S441 | D6S1017 | D4S2408 | D19S433 | D17S1301 | D1GATA113 | D18S853 | D20S482 | D14S1434 | D9S1122 | D2S1776 | D10S1435 | D5S2500 |
|--------|--------|----------|----------|----------|---------|----------|---------|---------|--------|--------|--------|--------|----------|----------|---------|---------|----------|---------|---------|----------|---------|
| 7      | 0.0036 | 0.2236   | 0.2364   | 0.0055   | 0.5145  | 0.0018   |
| 9      | 0.0036 | 0.2945   | 0.0019   | 0.0109   | 0.3582  | 0.2964   | 0.0655  | 0.0036  | 0.0913 | 0.0345 | 0.1073 | 0.0782 | 0.5318   | 0.0018   | 0.0036  | 0.1200  | 0.0036  |
| 10     | 0.0091 | 0.2546   | 0.0127   | 0.3473   | 0.0364  | 0.1545   | 0.0036  | 0.1782  | 0.1691 | 0.4036 | 0.1019 | 0.1636 | 0.1709   | 0.2818   | 0.1618  |
| 11     | 0.0018 | 0.2436   | 0.0109   | 0.0127   | 0.3473   | 0.0364   | 0.1545   | 0.0036  | 0.1782  | 0.1691 | 0.4036 | 0.1019 | 0.1636   | 0.1709   | 0.2818  |
| 12     | 0.3327 | 0.0018   | 0.0673   | 0.0182   | 0.0509  | 0.0982   | 0.0018  | 0.2109  | 0.2818  | 0.0182  | 0.0418 | 0.4582  | 0.2782   | 0.0582  | 0.0218  | 0.3182  | 0.3800  | 0.3473  |
| 13     | 0.0018 | 0.0036   | 0.0913   | 0.3545   | 0.0891  | 0.2073   | 0.5564  | 0.1891  | 0.0200  | 0.2836  | 0.5064  |
| 14     | 0.3745 | 0.0382   | 0.0109   | 0.2636   | 0.5309  | 0.3236   | 0.2927   | 0.2055  | 0.1200  | 0.0073  | 0.2545  | 0.5099  | 0.2309  | 0.3873  | 0.3709  | 0.0709  | 0.0036  | 0.1291  | 0.4073  |
| 15     | 0.3636 | 0.0018   | 0.2855   | 0.2200   | 0.3073  | 0.2855   | 0.0091   | 0.4182  | 0.0127  | 0.0509  | 0.0545  | 0.1855  | 0.0236  | 0.0036  | 0.0018  |
| 16     | 0.1345 | 0.1600   | 0.2582   | 0.0691   | 0.0418  | 0.0964   | 0.0018  | 0.1564  | 0.0218  | 0.0745  | 0.0745  | 0.0036  | 0.0036  |
| 17     | 0.3380 | 0.3957   | 0.1682   | 0.1227   | 0.0236  | 0.0291   | 0.0018  | 0.2873  | 0.2873  | 0.2873  | 0.2873  | 0.2873  | 0.2873  | 0.2873  | 0.2873  | 0.2873  | 0.2873  | 0.2873  | 0.2873  |
| 18     | 0.0327 | 0.0927   | 0.0218   | 0.0018   | 0.0036  | 0.0018   | 0.2327  | 0.0018  | 0.0018  | 0.0018  | 0.0018  | 0.0018  | 0.0018  | 0.0018  | 0.0018  | 0.0018  | 0.0018  | 0.0018  | 0.0018  |
| 19     | 0.0018 | 0.0010   | 0.0036   | 0.0018   | 0.0036  | 0.0018   | 0.0018  | 0.0018  |
| 20     | 0.0036 | 0.0036   | 0.0036   | 0.0036   | 0.0036  | 0.0036   | 0.0036  |
| 21     | 0.0036 | 0.0036   | 0.0036   | 0.0036   | 0.0036  | 0.0036   | 0.0036  |

PD: power of discrimination, PIC: polymorphism information content, PE: probability of exclusion, TPI: typical paternity index, HO: observed heterozygosity, HE: expected heterozygosity, p: probability values of exact tests for Hardy-Weinberg equilibrium.
Table 2 | Pairwise Fst and associated p values of 21 STR loci between Chinese Guanzhong Han population and 10 reference populations

| Population | index | Fst | p | p       | p       | p       | p       | p       | p       | p       | p       | p       | p       |
|------------|-------|-----|---|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|
| Bai        | -0.0018 | -0.0027 | -0.0028 | -0.0010 | 0.0004 | 0.0008 | 0.0061 | -0.0023 | 0.0034 | 0.0159 | 0.0015 | 0.0020 | -0.0020 | 0.0009 |
| Kazak      | 0.0017  | 0.0016  | 0.0014  | 0.0015  | -0.0024 | -0.0031 | 0.0019 | 0.0033 | 0.0165 | 0.0114 | 0.0079 | 0.0091 | 0.0094 | 0.0089 |
| Ningxia    | 0.0006  | 0.0002  | 0.0008  | -0.0019 | 0.0010 | -0.0009 | -0.0005 | -0.0008 | 0.0001 | -0.0008 | 0.0011 | -0.0009 | -0.0008 | 0.0007 |
| Han        | 0.9443  | 0.6452  | 0.4800  | 1.0000  | 0.3998 | 0.9834 | 0.9081 | 0.9961 | 0.9904 | 0.9414 | 0.3382 | 0.6833 | 0.9880 | 0.9853 |
| Russian    | -0.0026 | 0.0005  | 0.0036  | 0.0018  | 0.0020 | 0.0021 | 0.0018 | 0.0034 | 0.0042 | 0.0002 | 0.0001 | 0.0002 | 0.0002 | 0.0001 |
| Salar      | 1.0000  | 0.5992  | 0.8967  | 0.1230  | 1.0000 | 0.5992  | 0.8967  | 0.1230  | 1.0000 | 0.5992 | 0.8967  | 0.1230  | 1.0000 | 0.5992 |
| Tibetan    | 0.0036 | 0.0003  | 0.0015  | -0.0001 | 0.0008 | 0.0015  | -0.0001 | 0.0008 | 0.0015 | 0.0027 | 0.0010 | 0.0002 | 0.0001 | 0.0001 |
| Tuju        | -0.0102 | -0.0061 | -0.0034 | 0.0108  | 0.0036 | 0.0108  | 0.0036  | 0.0108 | 0.0036 | 0.0108  | 0.0036  | 0.0108 | 0.0036 | 0.0108 |
| Uigur      | 0.0010  | 0.0026  | 0.0025  | 0.0018  | 0.0025 | 0.0026  | 0.0025  | 0.0018  | 0.0025 | 0.0026  | 0.0025 | 0.0018 | 0.0025 | 0.0025 |
| Yi         | 0.0000  | 0.0000  | 0.0000  | 0.0000  | 0.0000 | 0.0000  | 0.0000  | 0.0000 | 0.0000 | 0.0000  | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| Mongolian  | 0.0012  | 0.0001  | 0.0011  | 0.0009  | 0.0034 | 0.0033  | 0.0026  | 0.0017 | 0.0062 | 0.0070  | 0.0022  | 0.0027 | 0.0005 | 0.0007 |
| Russian    | 0.9414  | 0.7498  | 0.5112  | 0.0411  | 0.0412 | 0.0287  | 0.0268  | 0.0269 | 1.0000 | 0.7359 | 0.0935  | 1.0000 | 0.0040 | 0.0676 |

The phylogenetic tree constructed by the software MEGA5 based on the Da distances was shown in Figure 2A. From the single-locus analyses, it was found that the Da distances among the 10 populations were highly significantly different (p < 0.0001). The phylogenetic tree based on the Da distances between the Guanzhong Han and 10 reference populations was shown in Figure 2B. The nodes of the tree, corresponding to the significance levels of the pairwise p values shown in Table 2, were shown in Figure 2B. The tree indicated that the Guanzhong Han population differed significantly from the other 10 previously published groups. The results indicated that the Guanzhong Han population was significantly different from the other 10 previously published groups at p < 0.0001. In addition, the results indicated that the Guanzhong Han population was significantly different from the other 10 previously published groups at p < 0.0001. However, the results indicated that the Guanzhong Han population was significantly different from the other 10 previously published groups at p < 0.0001.
Table 3 | The $D_A$ distances between Guanzhong Han population and other groups based on 10 STR loci

| Index | Ningxia Han | Tujia | Bai | Kazak | Tibetan | Mongolian | Uigur | Salar | Russian | Yi |
|-------|-------------|-------|-----|-------|---------|-----------|-------|-------|---------|----|
| $D_A$ | 0.0073      | 0.0077| 0.0091| 0.0126| 0.0133  | 0.0141   | 0.0153| 0.0264| 0.0281  | 0.0337|

Figure 2 | Phylogenetic tree for Guanzhong Han and 10 reference populations constructed by the software MEGA v5 based on $D_A$ distances (A) and by the software PHYLIP v3.6 based on allelic frequencies (B), respectively.

Figure 3 | Principal component analysis plot structured based on allelic frequencies of 21 STR loci in 11 populations.

Tibetan, Tujia and Bai groups shared the same clade; Yi, Russian, Salar and Mongolian groups were delineated in a branch; the remaining groups including Uigur and Kazak groups clustered together. In order to further confirm the phylogenetic relationship, the phylogenetic tree was also constructed using PHYLIP v3.6 based on the allelic frequencies of 21 STR loci and the result was shown in Figure 2B. The results obtained from two phylogenetic trees were extremely similar, and the only exception was Tibetan group. The exception may due to the different number of STR loci.

As shown in Figure 3, the PCA plot among 11 groups was obtained with the first two components to be 29.92% and 16.37%, respectively, which could explain 46.29% of the variance. The Guanzhong Han
population was observed to cluster closest with the Ningxia Han population, then with the Tujia and Bai groups, which is consistent with the results of phylogenetic trees above. The genetic evidence in our study showed that the Guanzhong Han population had closer relationship with Ningxia Han, Tujia and Bai populations than other 7 groups. The present result was basically consistent with the previous result of HLA loci as described by Shen et al. In order to further understand their genetic relationships and ancestry information, more genetic markers, such as SNPs and insertion/deletion polymorphisms should be used and analyzed in future.

Conclusions

In conclusion, we presented the genetic data of the Guanzhong Han population with 21 STR loci, and these STR loci showed high level of genetic polymorphisms and were suited for forensic application for the Guanzhong Han population. The population comparison showed the Guanzhong Han had a close genetic relationship with the Ningxia Han, Tujia and Bai populations among the populations tested.

1. Chu, J. Y. et al. Genetic relationship of populations in China. Proc. Natl. Acad. Sci. U S A. 95, 11763–11768 (1998).
2. Chen, J. et al. Genetic structure of the Han Chinese population revealed by genome-wide SNP variation. Am. J. Hum. Genet. 85, 775–785 (2009).
3. Qin, P. et al. A panel of ancestry informative markers to estimate and correct potential effects of population stratification in Han Chinese. Eur. J. Hum. Genet. 22, 248–253 (2014).
4. Xu, S. et al. Genomic dissection of population substructure of Han Chinese and its implication in association studies. Am. J. Hum. Genet. 85, 762–774 (2009).
5. Shen, C. M. et al. Allelic diversity and haplotype structure of HLA loci in the Chinese Han population living in the Guanzhong region of the Shaanxi province. Hum. Immunol. 71, 627–633 (2010).
6. Gao, Y. et al. Structural polymorphism analysis of Chinese Mongolian ethnic group revealed by a new STR panel: Genetic relationship to other groups. Electrophoresis 35, 2008–2013 (2014).
7. Shen, C. M. et al. Allelic polymorphic investigation of 21 autosomal short tandem repeat loci in a Chinese Bai ethnic group. Leg. Med. (Tokyo) 15, 109–113 (2013).
8. Yuan, J. Y. et al. Genetic profile characterization and population study of 21 autosomal STR in Chinese Kazak ethnic minority group. Electrophoresis 35, 503–510 (2014).
9. Wang, H. D. et al. Allelic diversity distributions of 21 new autosomal short tandem repeat loci in Chinese Ningxia Han population. Forensic Sci. Int. Genet. 7, 78–79 (2013).
10. Wang, H. D. et al. Allelic frequency distributions of 21 non-combined DNA index system STR loci in a Russian ethnic minority group from Inner Mongolia, China. J. Zhejiang Univ. Sci. B. 14, 533–540 (2013).
11. Zhu, B. F. et al. Genetic diversities of 21 non-CODIS autosomal STRs of a Chinese Tibetan ethnic minority group in Lhasa. Int. J. Legal Med. 125, 581–585 (2011).
12. Yuan, G. L. et al. Genetic data provided by 21 autosomal STR loci from Chinese Tujia ethnic group. Mol. Biol. Rep. 39, 10265–10271 (2012).
13. Deng, Y. J. et al. Polymorphic analysis of 21 new STR loci in Chinese Uigur group. Forensic Sci Int Genet. Forensic Sci. Int. Genet. 7, e97–e98 (2013).
14. Zhu, B. F. et al. Population genetics and forensic efficiency of twenty-one novel microsatellite loci of Chinese Yi ethnic group. Electrophoresis 34, 3345–3351 (2013).
15. Walsh, P. S., Metzger, D. A. & Higuchi, R. Chelex 100 as a medium for simple extraction of DNA for PCR-based typing from forensic material. Biotechniques 10, 506–513 (1991).
16. Teng, Y. et al. Genetic variation of new 21 autosomal short tandem repeat loci in a Chinese Salar ethnic group. Mol. Biol. Rep. 39, 1465–1470 (2012).
17. Tereba, A. Tools for Analysis of Population Statistics. profiles in DNA 2, 14–16 (1999).
18. Ota, T. DISPAN: genetic distance and phylogenetic analysis. (Pennsylvania State Univ. 1993).
19. Reich, D. E. et al. Linkage disequilibrium in the human genome. Nature 411, 199–204 (2001).