Biochem Genet (2011) 49:483–498
DOI 10.1007/s10528-011-9424-x

Dominant Contribution of Northern Chinese to the Paternal Genetic Structure of Chaoshanese in South China

Sheng-Ping Hu · Hui Li · Feng-Huan Zhang · Li-Qun Huang · Yan Lu

Received: 25 January 2010 / Accepted: 22 November 2010 / Published online: 1 March 2011

Abstract We investigated the Y chromosome of various Chinese populations to determine the patrilineal origin of the Chaoshanese population. Admixture analysis of six specific Y short tandem repeat (STR) loci in 6,292 individual samples taken from 51 populations, including Chaoshanese and Minnanese of our earlier studies, showed that over 85% of the Chaoshanese Y chromosomes were derived from the Central China Han (M_{RH}: 0.8614; M_{BE}: 1.1868 $\pm$ 0.2054), and a very small portion were from the southern aborigines. These results support a Central China Han origin of the Chaoshanese and additionally reveal that males from the Central China Han were the predominant contributor to the patrilineal genetics of the Chaoshanese. A phylogenetic tree and analysis of molecular variance signified a strong association between Y chromosomes of Chinese populations and their linguistic affiliations, revealing a coevolution of Y chromosome diversity and languages in East Asia.

Keywords Central China Han · Chaoshanese · Paternal genetic affinity · Southern indigenous population · Y-chromosomal short tandem repeat

S.-P. Hu (✉) · F.-H. Zhang
Molecular Biology Laboratory, Shantou University Medical College, 22 Xin Ling Road, Shantou 515031, Guangdong, China
e-mail: sphu@stu.edu.cn

S.-P. Hu
Forensic Genetics Laboratory, Shantou University Forensic Medical Service Center, Shantou 515031, Guangdong, China

H. Li · Y. Lu
MOE Key Lab of Contemporary Anthropology, School of Life Sciences, Fudan University, Shanghai 200433, China

L.-Q. Huang
Clinical Laboratory, First Affiliated Hospital, Shantou University Medical College, Shantou 515041, Guangdong, China

Introduction

The 56 ethnic populations of China can be classified into seven major groups based on their linguistic families or subfamilies: Altaic, Austro-Asiatic, Austronesian, Hmong-Mien, Daic (also called Baiyue in Chinese), Han (Sinitic), and Tibeto-Burman (Ethnologue Website http://www.ethnologue.com) (Wei and Wang 2000). Both the Han and the Tibeto-Burman belong to the Sino-Tibetan phylum (Martisoff 1991). The position of the Hmong-Mien and Daic in the language family tree, however, remains controversial. They are considered members of either the Sino-Tibetan family (Li 1977; Wei and Wang 2000) or of independent phyla (Benedict 1975) (http://www.ethnologue.com). Among the seven groups, Han is by far the largest in the world, with a population exceeding 1.18 billion, accounting for 90.6% of the population of China (2005 census www.stats.gov.cn/tjgb/rkpcgb/index.htm). According to Chinese history, the Han are descended from ancient Huaxia tribes that formed 4,000–5,000 years ago along the Yellow River valley in Central China and were dispersed throughout China during the past two millennia (Ge et al. 1997; Wei and Wang 2000). Tibeto-Burman speaking populations live mostly in China (provinces of Qinghai, Tibet, Sichuan, Yunnan, and Hunan) and in countries of South and Southeast Asia, whereas Altaic is spoken mainly in northern East Asia, including the upper northern part of China. The other four families (Daic, Austro-Asiatic, Austronesian, and Hmong-Mien) are native to southern China, Southeast Asia, and the Pacific Islands (Ethnologue Website) (Wang 1994). Among them, Daic is the largest and most widespread ethnic group in southern China, including Chaoshan (Song 1991; Huang 2002), with a population of 25.8 million (2,000 census), second only to the Han. It is known that the Daic are descendants of the Baiyue, an ancient aboriginal ethnic family that lived in the southern China coastal zone for at least 30,000 years before the Han arrived approximately 2,000 years ago (Song 1991; Wen et al. 2004a). Under the extensive influence of the Han, a large number of Baiyue became assimilated into the Han, and the rest migrated westward to become today's Daic populations (Li 2007). Daic is, therefore, significant in the study of the genetic structure of southern China populations (Li H et al. 2007, 2008).

The Chaoshanese, one of the three major Han ethnic groups in Guangdong province, are dwellers of the Chaoshan region, a littoral area located in southern China, with the South China Sea to the south and Fujian province to the east, across the strait from Taiwan. As recorded in Chaoshanese pedigrees and Chinese history, ancestors of the modern Chaoshanese originated from the Central China Han, who migrated to the Chaoshan area in continuous southward movements due to warfare and famine in the north (Ge et al. 1997; Wei and Wang 2000; Huang 2002). The first large-scale migration wave began in the Qin Dynasty around 214 BC ($\sim$2,200 years ago), when Emperor Qin sent battalions to unify the southland into one China. Over the years, in the blending of Han immigrants with the indigenous people, the Han assimilated aborigines and brought today's Chaoshanese into being. We have provided genetic evidence to support the Central China Han origin of the Chaoshanese by analyzing the allele frequency distribution of the human leukocyte antigen (HLA) complex A and B (Hu et al. 2007) and 15 autosomal-chromosomal short tandem repeat (STR) markers (Xu et al. 2009). We also showed that the

Chaoshanese are an admixture of the Central China Han and southern aboriginal natives (Xu et al. 2009).

Population admixture, one of the major themes in the study of human evolution, is being delineated for a number of populations, on different continents, by analyzing genetic variations using various genetic markers such as single nucleotide polymorphisms (SNP), mitochondrial DNA (mtDNA), and autosomal- and Y-chromosomal STRs (Y-STR) (Carvajal-Carmona et al. 2000; Benedetto et al. 2001; Gresham et al. 2001; Helgason et al. 2001; Sans et al. 2002; Bosch et al. 2003; Wen et al. 2004a, b; Kayser et al. 2008). Knowledge of population admixture enables an understanding of the historic, genetic, and evolutionary aspects of the relevant populations. Unequal contribution of male and female lineages from parental populations to filial generations is a common phenomenon seen in admixed populations (Carvajal-Carmona et al. 2000; Sans et al. 2002; Bosch et al. 2003; Wen et al. 2004a, b; Kayser et al. 2008). Moreover, the admixture ratio could be associated with the arrival time of immigrants and initial ethnic constitution of the parental populations in the regions. We previously described the admixture features of the Chaoshanese (Xu et al. 2009); however, the origin of male and female founders of this population has yet to be established. The Y-STR on the nonrecombining portion of the human Y chromosome has been proven useful in revealing the effects of male-influenced gene flow on the genetic composition, differentiation, and evolution of an admixed population (Quintana-Murci and Fellous 2001), and has been used in studies of genetic relationships among populations, population origin, diversity, admixture, migration, and evolution (Su et al. 1999; Carvajal-Carmona et al. 2000; Benedetto et al. 2001; Gresham et al. 2001; Sans et al. 2002; Bosch et al. 2003). In this study, we analyzed our Y-STR data (Hu 2006a, b), along with data collected from other Chinese populations of various linguistic families, to address the paternal genetic structure of the Chaoshanese and their relatedness to other Chinese populations in the male lineage.

Materials and Methods

Population Samples

The Y-STR data of Chaoshanese and Minnanese were derived from our previous studies (Hu 2006a, b), and those for the other Chinese populations were collected from peer-reviewed journals (Table 1). We used the following selection criteria: (1) population data published before 1 July 2008; (2) sample size of no less than 30; (3) populations with clearly stated ethnicity or geographic locations; and (4) samples with complete haplotype data for six specific Y-STR loci (*DYS19*, *DYS389I*, *DYS390*, *DYS391*, *DYS392*, and *DYS393*). Among the populations initially recruited by these criteria, several had more than one dataset from the same geographic or ethic group. The pairwise F_{ST} and associated P values of these datasets were first estimated with the haplotype frequencies using Arlequin 3.1 software (Excoffier et al. 2006). When the pairwise F_{ST} values were statistically insignificant

Table 1 Source of data for 51 Chinese populations

Geographic group	Population	Location	Linguistic family	References
Northern	Northeast Han	Northeast	Sinitic	Yang et al. (2006), Ba et al. (2007)
	Liaoning Han	Liaoning	Sinitic	Wang and Sawaguchi (2006)
	Ningxia Han	Ningxia	Sinitic	Zhu et al. (2006a)
	Tianjin Han	Tianjin	Sinitic	Kuang et al. (2005)
	Henan Han	Hennan	Sinitic	Feng et al. (2005)
	Shandong Han	Shandong	Sinitic	Yan et al. (2007)
	Beijing Han	Beijing	Sinitic	Kwak et al. (2005)
	Hui-1	Ningxia	Sinitic	Guo et al. (2008)
	Hui-2	Ningxia	Sinitic	Zhu YS et al. (2007)
	Korean	Jilin	Altaic	Zhang et al. (2007)
	Salar	Qinghai	Altaic	Zhu BF et al. (2007)
	Mongolian	Inner Mongolia	Altaic	Zhu et al. (2005a)
	Uigur	Xinjiang	Altaic	Zhu et al. (2005b)
	Dongxiang	Gansu	Altaic	Yang et al. (2005)
	Tibetan-1	Qinghai	Tibeto-Burman	Zhu et al. (2008)
Southern	Hunan Han	Hunan	Sinitic	Chen et al. (2005)
	Sichuan Han	Sichuan	Sinitic	Hidding and Schmitt (2000), Zhang et al. (2008)
	Yunnan Han	Yunnan	Sinitic	Zhang XH et al. (2006)
	Zhejiang Han	Zhejiang	Sinitic	Wu et al. (2005)
	Hong Kong Chinese	Hong Kong	Sinitic	Yeung et al. (2006)
	Taiwanese	Taiwan	Sinitic	Huang et al. (2008)
	Chaoshanese	Chaoshan	Sinitic	Hu (2006a)
	Minnanese	Minnan	Sinitic	Hu (2006b)
	Singapore Chinese	Singapore	Sinitic	Tang et al. (2006); Yong et al. (2006)
	Malaysia Chinese	Malaysia	Sinitic	Chang et al. (2007)
	Tibetan-2	Tibet	Tibeto-Burman	Zhang QX et al. (2006)
	Tibetan-3	Tibet	Tibeto-Burman	Kang et al. (2007)
	Tibetan-4	Tibet	Tibeto-Burman	Zhu et al. (2006c)
	Naxi	Yunnan	Tibeto-Burman	Xin et al. (2008)
	Yi	Yunnan	Tibeto-Burman	Zhu et al. (2006b)
	Tujia	Chongqing	Tibeto-Burman	Shi et al. (2008)
	Yao	Guangxi	Hmong-Mien	Liu (2004)
	Bolyu	Guangxi	Austro-Asiatic	Li H et al. (2008)
	Blue-Gelao	Guangxi	Daic	Li H et al. (2008)
	Lachi	Yunnan	Daic	Li H et al. (2008)
	Mollao	Guizhou	Daic	Li H et al. (2008)
	Red-Gelao	Guizhou	Daic	Li H et al. (2008)
	Hlai-Qi	Hainan	Daic	Li H et al. (2008)

Table 1 continued

Geographic group	Population	Location	Linguistic family	References
	Buyang	Yunnan	Daic	Li H et al. (2008)
	Cun	Hainan	Daic	Li H et al. (2008)
	Man-Caolan	Guangxi	Daic	Li H et al. (2008)
	Lingao	Hainan	Daic	Li H et al. (2008)
	E	Guangxi	Daic	Li H et al. (2008)
	Sui	Guangxi	Daic	Li H et al. (2008)
	Mak	Guizhou	Daic	Li H et al. (2008)
	Mulam	Guangxi	Daic	Li H et al. (2008)
	Maonan	Guangxi	Daic	Li H et al. (2008)
	Biao	Guangdong	Daic	Li H et al. (2008)
	Then	Guizhou	Daic	Li H et al. (2008)
	Danga	Hainan	Daic	Li H et al. (2008)
	DornQdayc	Shanghai	Daic	Li H et al. (2008)

($P > 0.05$), the datasets were combined. Otherwise, each dataset was used independently.

Accordingly, the current study included 6,292 individual samples from 51 populations with 1,697 haplotypes. These populations were divided into 15 northern and 36 southern populations with regard to the geographic locations separated at 30° north latitude (roughly along the Yangtze River) as previously described (Xiao et al. 2000). Alternatively, populations were grouped according to linguistic families into 19 Sinitic (including 7 Northern Han, 10 Southern Han, and 2 Hui), 7 Tibeto-Burman, 5 Altaic, 1 Hmong-Mien, 1 Austro-Asiatic, and 18 Daic (Table 1). It should be noted that the region we refer to as Central China is geographically located in northern China. We use the term "Central" to remain consistent with our previous studies.

Statistical Analysis

To quantify genetic relatedness among the studied populations, we calculated Nei's genetic distance *D* (Nei 1987) from Y-STR haplotype frequencies. Neighbor-joining trees (Saitou and Nei 1987) based on those genetic distances were constructed using programs in the Phylip 3.6 package (http://evolution.genetics.washington.edu/phylip.html) (Felsenstein 2004). Robustness of the tree was assessed by 1,000 bootstrap iterations (Felsenstein 1985). Population relationships were examined by principal component (PC) analysis based on Y-STR haplotype frequencies using SPSS 13.0 for Windows (SPSS Inc.). The gradient distribution of the PC values relative to the geographic location of each corresponding Han population was mapped using Surfer 8.0 (www.goldensoftware.com), in which derived PC1 and PC2 values served as height values. Population differentiation was evaluated by analysis of molecular variance (AMOVA) based on Y-STR haplotype frequencies and 10,000 permutations (Nei 1987; Slatkin 1995) using Arlequin 3.1 software

(Excoffier et al. 2006). The level of admixture of the Central China Han and southern natives in the Chaoshanese was estimated by Admix 2.0 (http://web.unife.it/progetti/genetica/Isabelle/admix2_0.html) (Dupanloup and Bertorelle 2001) and Leadmix (http://www.zoo.cam.ac.uk/ioz/software.htm) (Wang 2003), using M_{BE} (Bertorelle and Excoffier 1998) and M_{RH} statistics (Roberts and Hiorns 1965). In this analysis, the haplotype frequency of the Henan Han was used to represent the Central China Han. The average (arithmetic mean) haplotype frequencies of two southern native populations, Hlai and Cun (Li H et al. 2008), were taken for that population, as described (Wen et al. 2004a). We assumed the potential admixture started 2,300 years ago, when the Qin army (Central China Han) first entered the Chaoshan area (Ge et al. 1997; Wei and Wang 2000; Huang 2002).

Results

Genetic Relationship Among the Studied Populations

All populations finally selected were included in the neighbor-joining phylogenetic analysis (Fig. 1) and the PC analysis (Fig. 2). All studied populations clustered on the basis of their linguistic families irrespective of the geographic division. This phenomenon was even more evident in the PC analysis. The cumulative contribution of PC1 and PC2 accounted for 41.71% of the total difference. In the PC plot, the Han and Daic each fell into their respective linguistic families and were distinctly separated by both PC1 and PC2, with the Daic clustering together in the upper left part of the plot and the Han in the lower right. Within the Han cluster, the distinction between Southern and Northern Hans was noticeable by PC1, but less evident. It is interesting that the Henan Han lay closer to the Southern Han, consistent with what we have previously observed by both HLA (Hu et al. 2007) and autosomal STR analyses (Xu et al. 2009). The other populations formed a loose gathering in the lower left corner of the plot, with less distinct boundaries between their linguistic families. This loose gathering was separated from Daic and Hans by PC2 and PC1, respectively, and the lack of distinction between the linguistic families may reflect the smaller number of populations sampled in each of those linguistic families in this study. These results indicate a highly coherent paternal genetic structure of Chinese populations that is associated with the linguistic family.

The genetic affinity between the Chaoshanese and the other 16 Han populations was further visualized in a neighbor-joining tree (Fig. 3) and a geographic map of the PC1 and PC2 values (Fig. 4). The tree for Han populations shows results similar to those in the first tree. The Han populations clustered together as one linguistic family, and separation was evident between the Southern and Northern Hans within the family. The Chaoshanese clustered in general with the Southern Han but appeared, among the Southern Han, to be more closely related to the Minnanese, Taiwanese, Singapore Chinese, and Malaysia Chinese. The D values between the Chaoshanese and these four populations were 0.000006, 0.000008, 0.000008, and 0.00001, respectively. These populations also fall into the same grade in the PC maps (Fig. 4). Among the Northern Han populations, the Henan Han appeared to be

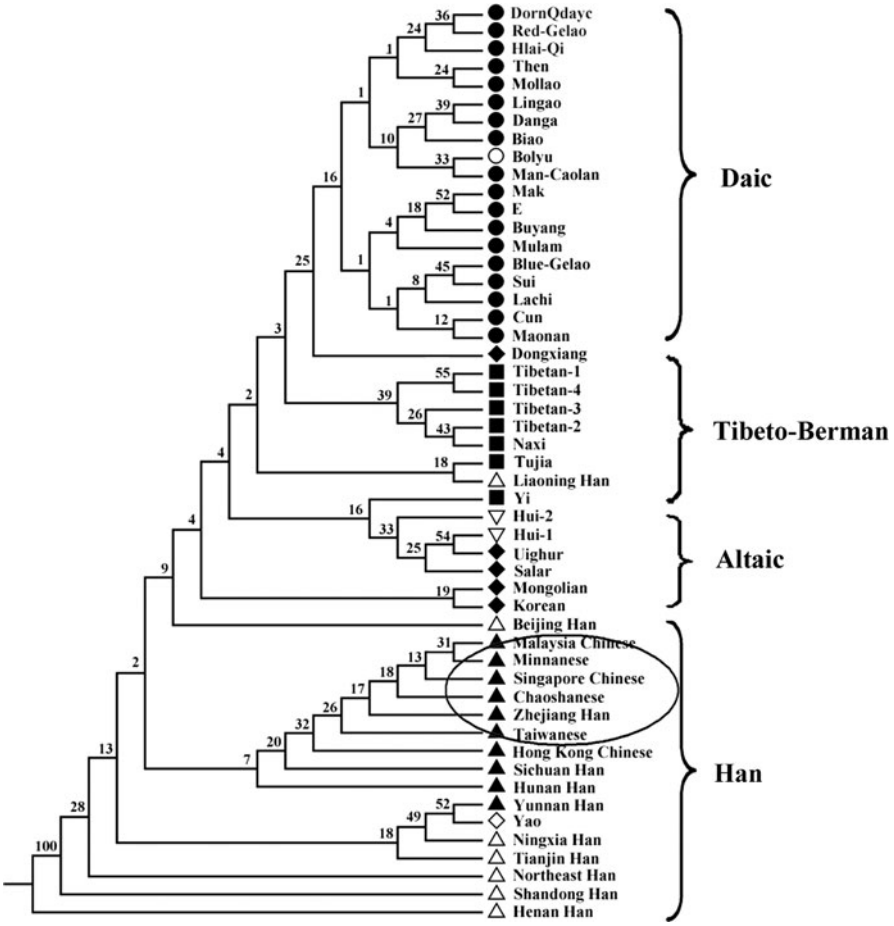

Fig. 1 Phylogenetic relationships of the 51 populations in the present study, based on Y-STR haplotype frequencies. Neighbor-joining tree calculated from Nei's genetic distance *D* (Nei 1987). The numbers above the nodes are bootstrap values (%) estimated from 1,000 replicates. Linguistic groups are indicated by geometric shapes: *open triangle* Northern Han, *filled triangle* Southern Han, *open inverted triangle* Hui, *filled square* Tibeto-Burman, *filled diamond* Altaic, *open diamond* Hmong-Mien, *filled circle* Daic, *open circle* Austro-Asiatic

closest to the Southern Han populations, with a *D* value of 0.000009, the smallest value among the distances between the Chaoshanese and other Northern Han populations. This indicates a close genetic relationship between the Central China Han and the Southern Han populations to a certain extent. Moreover, the Chaoshan-Minnan grade was close to that of the Central China Han populations (Fig. 4).

Analysis of Molecular Variance (AMOVA)

Populations of the same linguistic family clustered together in our phylogenetic and PC analyses (Figs. 1, 2). If this clustering has a genetic basis, the variance among

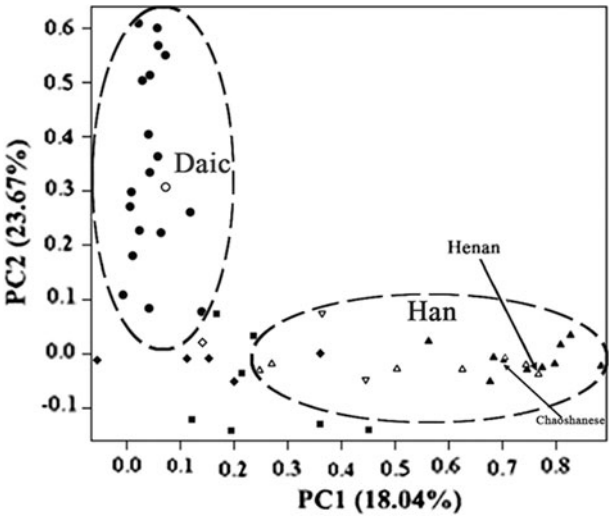

Fig. 2 Principal component plot of the 51 populations in the present study. Each population is represented by the symbol for its linguistic group, as listed in Fig. 1

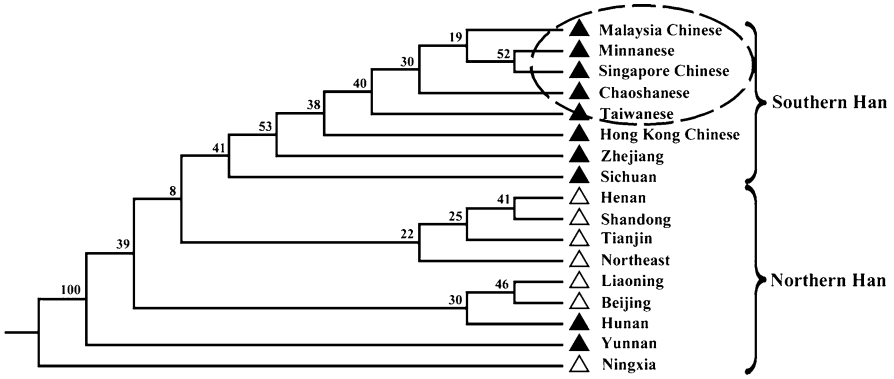

Fig. 3 Phylogenetic tree of 17 Han populations, based on the haplotype frequencies of six specific Y-STR loci. The numbers in the nodes are bootstrap values (%) estimated from 1,000 replicates. *Open triangles* indicate Northern Han linguistic group; *solid triangles* indicate Southern Han

linguistic groups should be greater, whereas the variance among geographic groups should be smaller in the AMOVA analysis. We divided the study populations into southern and northern groups (geographic groups) and also grouped them by their linguistic families (linguistic groups) for AMOVA analysis to determine the geographic and linguistic partitioning of Y-STR diversity (Table 2). As anticipated, Y-chromosome differentiation was far greater among linguistic groups (0.56%; $F_{CT} = 0.00563$, $P = 0.000$) than between the geographic groups (0.08%; $F_{CT} = 0.00077$, $P = 0.016$). The F_{CT} among linguistic groups was much higher than that between geographic groups, though F_{CT} statistics of both groups were

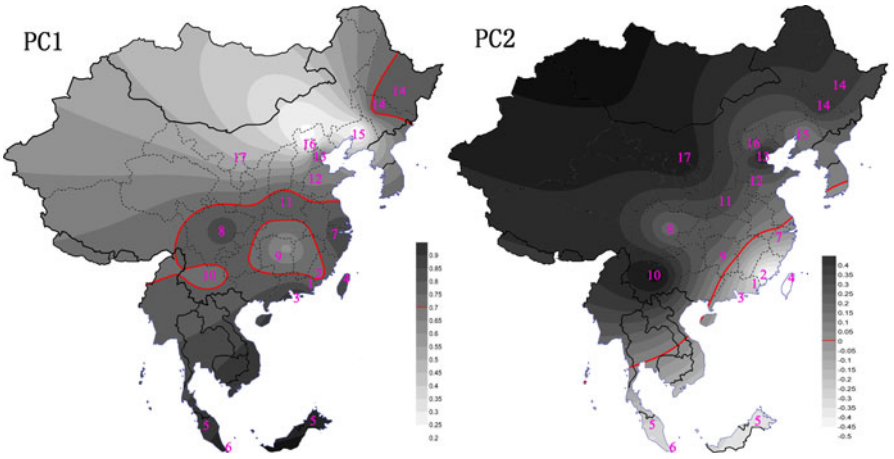

Fig. 4 Principal component maps of Y-STR from 17 Chinese Han populations. Scales indicate the gradient distribution of the values plotted in Fig. 2 for PC1 (*left*) and PC2 (*right*) relative to the geographic location of the numbered populations: 1. Chaoshanese; 2. Minnanese; 3. Hong Kong Chinese; 4. Taiwanese; 5. Malaysia Chinese; 6. Singapore Chinese; 7. Zhejiang; 8. Sichuan; 9. Hunan; 10. Yunnan; 11. Henan; 12. Shandong; 13. Tianjin; 14. Northeast; 15. Liaoning; 16. Beijing; 17. Ningxia

Table 2 AMOVA of 51 Chinese populations

Group	Source of variation		
	Among groups (F_{CT}, P)	Among populations within groups (F_{SC}, P)	Within populations (F_{ST}, P)
Geographic	0.08 (0.00077, 0.016)	0.84 (0.00839, 0.000)	99.08 (0.00916, 0.000)
Linguistic	0.56 (0.00563, 0.000)	0.55 (0.00549, 0.000)	98.89 (0.01110, 0.000)

significant ($P < 0.05$), suggesting a more extensive among-groups differentiation when populations were divided by linguistic families. In contrast, the variance among populations within groups was quite the opposite, with 0.84% ($F_{SC} = 0.00839$, $P = 0.000$) for populations within the geographic groups versus 0.55% ($F_{SC} = 0.00549$, $P = 0.000$) for those within linguistic groups. These results clearly indicate that the paternal genetic linkage among the Chinese populations was associated with their linguistic affiliations more than with their geographic locations.

Admixture Analysis of the Chaoshanese

Historical records and our previous autosomal STR data suggest two potential source populations for Chaoshanese Y chromosomes: the Central China Han and the Daic (Ge et al. 1997; Wei and Wang 2000; Xu et al. 2009). We considered the Henan Han as the parental population representative of the Central China Han because Henan covers ancient Central China (Hu et al. 2007; Xu et al. 2009). Hlai and Cun, two populations that we used as the southern parental population, are

Table 3 Admixture analysis of the Chaoshanese population

Method	Parental population	
	Central China Han	Southern indigenous
Leadmix (M_{RH})	0.8614	0.1386
Admix (M_{BE})	1.1868 ± 0.2054	−0.1868 ± 0.2054

believed to be genetically closest to the ancestral Daic (Li H et al. 2008). These two populations have been living on Hainan Island for generations (Li D et al. 2008). The Qiongzhou Strait, which lies between the island and the mainland, was a formidable obstacle to communications until recently, isolating Hainan aborigines from the influence of exotic ethnic population relocation and admixture (Li D et al. 2008). The two populations thus could serve as representatives of southern natives (Wen et al. 2004a). Estimates of the admixture coefficient from the two programs were highly consistent (Table 3). Chaoshanese showed a dominant proportion of the Central China Han contribution (M_{RH}: 0.8614; M_{BE}: 1.1868 ± 0.2054), but only a small part from southern natives (M_{RH}: 0.1386; M_{BE}: −0.1868 ± 0.2054). These results indicate that the Chaoshanese paternal gene pool consists almost exclusively of Central China Han with little input of southern indigenous origin.

Discussion

In this study, we extended our prior research to determine the origin of the male founder of the Chaoshanese and the paternal affinity of this population with other Chinese ethnic groups. By estimating the relative contributions of the two parental populations in Chaoshanese, using two statistics (Roberts and Hiorns 1965; Bertorelle and Excoffier 1998), we for the first time showed that the Y chromosome of the Chaoshanese was derived overwhelmingly from the Central China Han, with very little contribution from southern natives (Table 3). These results confirm our prior demonstration of the Central China Han origin of the Chaoshanese (Xu et al. 2009) and further show that males from Central China Han were the predominant contributors to the paternal gene pool of Chaoshanese, as judged by the high patrilineal similarity shared by the Chaoshanese and their parental Central China Han.

High paternal resemblance to Northern Hans (males from Northern Hans are the primary contributors to the paternal gene pool of the filial generation), along with considerable maternal differences from either parental population (all parental populations contribute substantially to the maternal gene pool of filial generation), has been previously documented for other Southern Han populations (Wen et al. 2004a). The mechanism underlying this pattern has been explained as a strong sex-biased population admixture due to "directional mating" (Merriwether et al. 1997). The Chaoshanese belong to the Southern Han (Figs. 1, 2, 3) (Huang 2002; Hu et al. 2007; Xu et al. 2009). We therefore assume that the same admixture process may have been followed by this population. This is likely to be true as the Chaoshanese,

like other Southern Hans (Wen et al. 2004a), showed approximately equal proportions of maternal genetic components from both the Central China Han and the southern natives (unpublished observations). It has been previously suggested for different world populations that directional mating is governed strongly by several factors, including sex ratio of migration, mate choice, and linguistic, and social attributes (Bamshad et al. 1998; Seielstad et al. 1998; Pérez-Lezaun et al. 1999; Carvajal-Carmona et al. 2000; Benedetto et al. 2001; Parra et al. 2001; Sans et al. 2002; Bosch et al. 2003; Wen et al. 2004a, b; Kayser et al. 2008). The admixture genetics presented here is consistent with the historical and demographic information available on the Chaoshanese. Regarding migration in Chaoshanese history, southward expansion of the Central China Han to Chaoshan largely involved expedition forces primarily consisting of young or middle-aged male soldiers (Huang 2002; Chen 2006; Hu et al. 2007; Xu et al. 2009). Such an asymmetrically high male proportion in migrants would have inevitably affected the mate choice, leaving indigenous women as the primary mating targets (Fan 2000). From a cultural practice aspect, the Han migration southward followed a demic diffusion model, in which the mass movement of Han people spread the Han culture and language (Wen et al. 2004a). Possessing advanced technologies and culture, the Hans would undoubtedly exert a dominant and extensive influence on indigenous people (Ge et al. 1997). Specifically, a matrimonial pattern of patrilocal residence (wives move into the husband's household) and patrilineal descent (clans are inherited through the paternal line) was strictly practiced in old-time China (Chen 2006). This would have greatly favored the incorporation and retention of paternal Central China Han genes in the ancestral Chaoshanese. Indeed, immigrant soldiers were encouraged to marry indigenous women when they settled in Chaoshan after wars (Chen 2006). Intermarriage was significant in easing conflict between military immigrants and indigenous people, and in speeding up cultural exchange and national integration (Fan 2000). It is worth mentioning that most of the cultural practices in today's Chaoshanese reflect a Central China Han origin (Chen 2006).

A distinction between northern and southern Chinese populations, separated at 30° north latitude, approximately by the Yangtze River (Xiao et al. 2000), has been previously observed by analysis of genetic markers other than Y variations, as well as somatometric and nonmetric features in our studies (Hu et al. 2007; Xu et al. 2009; unpublished observation) and others (Zhao and Lee 1989; Du et al. 1992; Chu et al. 1998; Xiao et al. 2000; Yao et al. 2002; Wen et al. 2004a). Contrary to this observation, however, our phylogenetic and PC analyses (Figs. 1, 2) on Y variations showed that all the populations clustered essentially by linguistic family, regardless of their geographic locations and distances. In other words, on the paternal side, Chinese populations tended to group with their linguistic neighbors, not their geographic neighbors. This is particularly obvious for those linguistic groups that include populations from geographically divided northern and southern groups, such as the Tibeto-Burman and Han. We further showed that language-based clustering was evident even within a linguistic family. The Tibeto-Burman and Han are both in the Sino-Tibetan family, but each formed its own assembly (Fig. 1). Chaoshanese, Minnanese, Taiwanese (Lin et al. 2001), Malaysia Chinese, and Singapore Chinese, though geographically far apart, clustered together more closely among the

Southern Han populations (Figs. 1, 4). These five ethnic groups speak the Min dialect, one of the 10 major dialects within the Han family (Gan et al. 2008), and share common ancestors (Huang 2002; Tang 2002; Lin and Qiu 2005; Zhang 2006; Hu et al. 2007; Xu et al. 2009). Our results indicate that the paternal genetic relationships among Chinese populations correlated more strongly with linguistic rather than geographic relationships, in agreement with prior observations in worldwide populations that linguistic distances present a higher correlation with Y chromosome than with mtDNA markers (Cavalli-Sforza et al. 1992; Chen et al. 1995; Poloni et al. 1997). It is known that the matrimonial pattern has a strong impact on the genetic structure of a population. In a society centered with patrilocality and patrilineality, men possess both cultural and political dominance. This would make society members less motivated to change their life style, but rather maintain their customs for generations. As we mentioned above, patrilocality and patrilineality are indigenous to traditional Chinese culture and are still being practiced in the countryside of China. The same mechanism that keeps coevolution of Y chromosome and language would therefore also operate in Chinese populations.

In summary, we have for the first time analyzed the paternal genetic structure of the Chaoshanese and their affinity with other Chinese populations along the male lineage. Our results confirm the Central China Han origin of the Chaoshanese and further reveal that the Y chromosomes of Chaoshanese are derived predominantly from the Central China Han with little contribution from southern natives. Moreover, on the paternal side, Chinese populations tended to cluster based on their linguistic affiliations rather than their geographic distribution. The geographic boundary between northern and southern Chinese populations, as observed in other analyses, was no longer distinct.

Acknowledgments This work was supported by research grants from the Li Ka-Shing Foundation, Hong Kong; Cambridge University, UK; the Natural Science Foundation of Guangdong Province, China (No. 000819); the Shantou University Research and Development Foundation (No. L00007), and the Shantou Key Research Project, Shantou Science and Technology Bureau (No. 2005116).

References

Ba HJ, Lin ZQ, Li S (2007) Polymorphisms of 11 Y-STR loci in northeast China Han. J Forensic Med 23:206–209

Bamshad MJ, Watkins WS, Dixon ME, Jorde LB, Rao BB, Naidu JM, Prasad BVR (1998) Female gene flow stratifies Hindu castes. Nature 395:651–652

Benedetto GD, Ergüven A, Stenico M, Castrfi L, Bertorelle G, Togan L, Barbujani G (2001) DNA diversity and population admixture in Anatolia. Am J Phys Anthropol 115:144–156

Benedict PK (1975) Austro-Thai language and culture, with a glossary of roots. HRAF Press, New Haven, Conn

Bertorelle G, Excoffier L (1998) Inferring admixture proportions from molecular data. Mol Biol Evol 15:1298–1311

Bosch E, Calafell F, Rosser ZH, Norby S, Lynnerup N, Hurles ME, Jobling MA (2003) High level of male-biased Scandinavian admixture in Greenlandic Inuit shown by Y-chromosomal analysis. Hum Genet 112:353–363

Carvajal-Carmona LG, Soto LD, Pineda N, Ortiz-Barrientos D, Duque C, Qspina-Duque J, McCarthy M, Montoya P, Alvarez VM, Bedoya G, Ruiz-Linares A (2000) Strong Amerind/white sex bias and a

possible sephardic contribution among the founders of a population in northwest Colombia. Am J Hum Genet 67:287–1295
Cavalli-Sforza LL, Minch E, Mountain JL (1992) Coevolution of genes and languages revisited. Proc Natl Acad Sci 89:5620–5624
Chang YM, Perumal R, Keat PY, Kuehn DLC (2007) Haplotype diversity of 16 Y-chromosomal STRs in three main ethnic populations (Malays, Chinese and Indians) in Malaysia. Forensic Sci Int 167:70–76
Chen XX (2006) The origin of ancestors in Chaoshan, 1st edn. Guangdong People's Publishing House, Guangzhou
Chen J, Sokal RR, Ruhlen M (1995) Worldwide analysis of genetic and linguistic relationships of human populations. Hum Biol 67:595–612
Chen SQ, Chen HJ, Zeng X, Li Q, Zhu ZL, Nie SL (2005) Polymorphisms of 12 Y-chromosome STR loci in Han population in Hunan. Chin J Forensic Med 20:174–176
Chu JY, Huang W, Kuang SQ, Wang JM, Xu JJ, Chu ZT, Yang ZQ, Lin KQ, Li P, Wu M, Geng ZC, Tan CC, Du RF, Jin L (1998) Genetic relationship of populations in China. Proc Natl Acad Sci 95:11763–11768
Du RF, Yuan YD, Hwang J, Mountain J, Cavalli-Sforze LL (1992) Chinese surnames and the genetic differences between north and south China. J Chin Ling, monograph series 5
Dupanloup I, Bertorelle G (2001) Inferring admixture proportions from molecular data: extension to any number or parental populations. Mol Biol Evol 18:672–675
Excoffier L, Guillaume L, Schneider S (2006) Computational and Molecular Population Genetics Lab (CMPG). University of Bern, Switzerland
Fan YC (2000) On the effects of ancient Chinese military immigration. J Guangxi Normal University (Philosophy and Social Science Edition) 36:81–84
Felsenstein J (1985) Confidence limits on phylogenies: an approach using the bootstrap. Evolution 39:783–791
Felsenstein J (2004) Phylip (Phylogeny Inference Package) version 3.6. University of Washington, Department of Genome Sciences, Seattle
Feng CJ, Xiang ZD, Shen CB (2005) Polymorphisms of 12 Y-chromosome STR loci in Han population in Henan. Forensic Sci Technol (China) 177:23–28
Gan RJ, Pan SL, Mustavich LF, Qin ZD, Cai XY, Qian J, Liu CW, Peng JH, Li SL, Xu JS, Jin L, Li H (2008) Pinghua population as an exception of Han Chinese's coherent genetic structure. J Hum Genet 53:303–313
Ge JX, Wu SD, Chao SJ (1997) The migration history of China (in Chinese). Fujian People's Publishing House, Fuzhou
Gresham D, Morar B, Underhill PA, Passarino G, Lin AA, Wise C, Angelicheva D, Calafell F, Oefner PJ, Shen PD, Tournev L, Pablo RD, Kuĉinskas V, Perez-Leazun A, Marushiakova E, Popov V, Kalaydjieva L (2001) Origins and divergence of the Roma (gypsies). Am J Hum Genet 69:1314–1331
Guo H, Yan JW, Jiao ZP, Tang H, Zhang QX, Zhao L, Hu N, Li HF, Liu YC (2008) Genetic polymorphisms for 17 Y-chromosomal STRs haplotypes in Chinese Hui population. Leg Med 10:163–169
Helgason A, Hickey E, Goodacre S, Bosnes V, Stefánsson K, Ward R, Sykes B (2001) mtDNA and the islands of the North Atlantic: estimating the proportions of Norse and Gaelic ancestry. Am J Hum Genet 68:723–737
Hidding M, Schmitt C (2000) Haplotype frequencies and population data of nine Y-chromosomal STR polymorphisms in a German and a Chinese population. Forensic Sci Int 113:47–53
Hu SP (2006a) Polymorphism of Y-chromosomal STR haplotypes in the Chaoshan Han Chinese in South China. Forensic Sci Int 158:80–85
Hu SP (2006b) Genetic Polymorphism of 12 Y-chromosomal STR loci in the Minnan Han Chinese in Southeast China. Forensic Sci Int 159:77–82
Hu SP, Luan JA, Li B, Chen JX, Cai KL, Huang LQ, Xu XY (2007) Genetic link between Chaoshan and other Chinese Han populations: evidence from HLA-A and HLA-B allele frequency distribution. Am J Phys Anthropol 132:140–150
Huang T (2002) Headstream of Chaoshan culture. Guangdong Higher Education Press, Guangzhou
Huang TY, Hsu YT, Li JM, Chung JH, Shun CT (2008) Polymorphism of 17 Y-STR loci in Taiwan population. Forensic Sci Int 174:249–254

Kang LL, Liu K, Ma YM (2007) Y chromosome STR haplotypes of Tibetan Living Tibet Lassa. Forensic Sci Int 172:79–83

Kayser M, Lao O, Saar K, Brauer S, Wang XY, Nürnber P, Trent RJ, Stoneking M (2008) Genome-wide analysis indicates more Asian than Melanesian ancestry of Polynesians. Am J Hum Genet 82:194–198

Kuang JZ, Zhu W, Nie TG, Liu Y, Liu MN, Wang YJ (2005) Polymophisms of 12 Y-STR loci in Han population in Tianjin. Forensic Sci Technol (China) 175:19–26

Kwak KD, Jin HJ, Shin DJ, Kim JM, Roewer L, Krawezak M, Tyler-Smith C, Kim W (2005) Y-chromosomal STR haplotypes and their applications to forensic and population studies in east Asia. Int Leg Med 119:195–201

Li D, Li H, Ou CY, Lu Y, Sun YT, Yang B, Qin ZD, Zhou ZJ, Li SL, Jin L (2008) Paternal genetic structure of Hainan aborigines isolated at the entrance to East Asia. PLoS One 3:e2168

Li FK (1977) A handbook of comparative Tai. University Press of Hawaii, Honolulu

Li H (2007) Abscondence of Min-Yue ethnic group revealed by molecular anthropology (in Chinese). J Guangxi University for Nationalities (Philosophy and Social Science Edition) 29:42–47

Li H, Cai XY, Winograd-Cort ER, Wen B, Cheng X, Qin ZD, Liu WH, Liu YF, Pan SL, Qian J, Tan CC, Jin L (2007) Mitochondrial DNA diversity and population differentiation in Southern East Asia. Am J Phys Anthropol 184:481–488

Li H, Wen B, Chen SJ, Su B, Pramoonjago P, Liu YF, Pan SL, Qin ZD, Liu WH, Cheng X, Yang NN, Li X, Tran D, Lu D, Hsu MT, Deka RJ, Marzuki S, Tan CC, Jin L (2008) Paternal genetic affinity between western Austronesians and Daic populations. BMC Evol Biol 8:146

Lin GP, Qiu JD (2005) The migration history of Fujian. Fangzhi Publishing House, Beijing

Lin M, Chu CC, Chang SL, Lee HL, Loo JH, Akaza T, Juji T, Ohashi J, Tokunaga K (2001) The origin of Minnan and Hakka, the so-called "Taiwanese", inferred by HLA study. Tissue Antigens 57:192–199

Liu Y (2004) Polymorphism of Y-STR loci in Yao ethnic group from Du'an of Guangxi. Master Thesis, part 2, Forensic Sci, University of Sichuan, pp 36–54

Martisoff JA (1991) Sino-Tibetan linguistics: present state and future prospects. Annu Rev Anthropol 20:469–504

Merriwether DA, Huston S, Iyengar S, Hamman R, Norris JM, Shetterly SM, Kamboh MI, Ferrell RE (1997) Mitochondrial versus nuclear admixture estimates demonstrate a past history of directional mating. Am J Phys Anthropol 102:153–159

Nei M (1987) Molecular evolutionary genetics. Columbia University Press, New York

Parra EJ, Kittles RA, Argyropoulos G, Pfaff CL, Hiester K, Bonilla C, Sylvester N, Parrish-Gause D, Garvey WT, Jin L, McKelgue PM, Kamboh ML, Ferrell RE, Pollltzer WS, Shrlver MD (2001) Ancestral proportions and admixture dynamics in geographically defined African Americans living in South Carolina. Am J Phys Anthropol 114:18–29

Pérez-Lezaun A, Calafell F, Comas D, Mateu E, Bosch E, Martinez-Arias R, Clarimón J, Fiori G, Luiselli D, Facchini F, Pettener D, Bertranpetit J (1999) Sex-specific migration patterns in Central Asian populations, revealed by analysis of Y-chromosome short tandem repeats and mtDNA. Am J Hum Genet 65:208–219

Poloni ES, Semino O, Passarino G, Santachiara-Benerecetti AS, Dupanloup L, Langaney A, Excoffier L (1997) Human genetic affinities for Y-chromosome p49a, f/*Taql* haplotypes show strong correspondence with linguistics. Am J Hum Genet 61:1015–1035

Quintana-Murci L, Fellous M (2001) The human Y chromosome: the biological role of a "functional wasteland". J Biomed Biotechnol 1:18–24

Roberts DF, Hiorns RW (1965) Methods of analysis of the genetic composition of a hybrid population. Hum Biol 37:38–43

Saitou N, Nei M (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol 4:406–425

Sans M, Welmer TA, Franco MHLP, Salzano FM, Bentancor N, Alvarez L, Blanchl NO, Chakraborty R (2002) Unequal contributions of male and female gene pools from parental populations in the African descendants of the city of Melo, Uruguay. Am J Phys Anthropol 118:33–44

Seielstad MT, Minch E, Cavalli-Sforza L (1998) Genetic evidence for a higher female migration rate in humans. Nat Genet 20:278–280

Shi MS, Bai R, Wan L, Yu X, Chang L (2008) Population genetics for Y-chromosomal STRs haplotypes of Chinese Tujia ethnic group. Forensic Sci Int Genet 2:e65–e68

Slatkin M (1995) A measure of population subdivision based on microsatellite allele frequencies. Genetics 139:457–462

Song SH (1991) Bai-Yue (in Chinese). Jilin Education Press, Changchun

Su B, Xiao JH, Underhill P, Deka R, Zhang WL, Akey J, Huang WH, Shen D, Lu D, Luo JC, Chu JY, Tan JZ, Shen PD, Davis R, Cavalli-Sforza L, Chakraborty R, Xiong MM, Du RF, Oefner P, Chen Z, Jin L (1999) Y-chromosome evidence for a northward migration of modern humans into Eastern Asia during the last Ice Age. Am J Hum Genet 65:1718–1724

Tang JSW, Wong HY, Syn CKC, Tan-Siew WF, Chow ST, Budowle B (2006) Population study of 11 Y-chromosomal STR loci in Singapore Chinese. Forensic Sci Int 158:65–71

Tang ZP (2002) Immigrants of the Central Plains into Fujian in the Early Tang Dynasty and the formation of Fujian and Taiwan Culture. J Xuchang Teach Coll 21:82–86

Wang J (2003) Maximum-likelihood estimation of admixture proportions from genetic data. Genetics 164:747–765

Wang XL, Sawaguchi T (2006) Analysis of Y-STR loci in a population sample from Northeast China. J Forensic Sci 51:195–198

Wang ZH (1994) History of nationalities in China (in Chinese). China Social Science Press, Beijing

Wei DC, Wang RL (2000) Migration and evolution history of Chinese ethnic groups (in Chinese). Hubei People's Publishing House, Wuhan

Wen B, Li H, Lu D, Song XF, Zhang F, He YG, Li F, Gao Y, Mao XY, Zhang L, Qian J, Tan JZ, Jin JZ, Huang WH, Deka RJ, Su B, Chakraborty R, Jin L (2004a) Genetic evidence supports demic diffusion of Han culture. Nature 431:302–305

Wen B, Xie XH, Gao S, Li H, Shi H, Song XF, Qian TZ, Xiao CJ, Jin JZ, Su B, Lu D, Chakraborty R, Jin L (2004b) Analyses of genetic structure of Tibeto-Burman populations reveals sex-biased admixture in Southern Tibeto-Burmans. Am J Hum Genet 74:856–865

Wu WW, Zheng XT, Pan LP, Hao HL, Fu T (2005) A study of polymorphisms of 16 Y-STR loci in Han population in Zhejiang. Forensic Sci Technol (China) 179:11–17

Xiao CJ, Cavalli-Sforza LL, Minch E, Du RF (2000) Principal component analysis of gene frequencies of Chinese populations. Sci China C 43:472–481

Xin N, Chen T, Yu B, Li SB (2008) 12 Y-STRs haplotypes in Chinese Naxi ethnic minority Group. Forensic Sci Int 174:244–248

Xu LN, Hu SP, Feng GY (2009) STR polymorphisms of the Henan population and investigation of the Central Plains Han origin of Chaoshanese. Biochem Genet 47:569–581

Yan JW, Tang H, Liu YC, Jing YT, Jiao ZP, Zhang QX, Gao JW, Shang LP, Guo H, Yu J (2007) Genetic polymorphisms of 17 Y-STRs haplotypes in Chinese Han population residing in Shandong province of China. Leg Med 9:196–202

Yang BQ, Gu M, Wang G, Li X, Liu Y, Yang W (2006) Population data for 11 Y-chromosome STRs in northeast China Han. Forensic Sci Int 164:65–71

Yang YJ, An LZ, Xu JJ, Zhang WH, Zhou RX, Wang XL, Xie XD (2005) Human Y-specific STR haplotypes in Dongxiang ethnic group from Northwest China. J Lanzhou University (Nature Science) 41:34–38

Yao YG, Kong QP, Bandelt HJ, Kivisild T, Zhang YP (2002) Phylogeographic differentiation of mitochondrial DNA in Han Chinese. Am J Hum Genet 70:635–651

Yeung SM, Wong LM, Cheung BKK, To KY (2006) Allele frequencies and haplotypes of 12 Y-STR loci for the local Chinese population in Hong Kong. Forensic Sci Int 162:55–63

Yong RYY, Lee LKH, Yap EPH (2006) Y-chromosome STR haplotype diversity in three ethnic populations in Singapore. Forensic Sci Int 159:244–257

Zhang HJ, Yun LB, Li YB, Zhang J, Wu J, Yan J, Hou YP (2008) Haplotype of 12 Y-STR loci of the PowerPlex Y-system in Sichuan Han ethnic group in west China. Forensic Sci Int 175:244–249

Zhang QX, Yan JW, Tang H, Jiao ZP, Liu YC (2006) Genetic polymorphisms of 17 Y-STRs haplotypes in Tibetan ethnic minority group of China. Leg Med 8:300–305

Zhang XH, Wu WW, Tang JX, Qian GL, Zhang XM (2006) Polymorphisms of 11 Y-chromosome STR loci and forensic application in Yunnan Han population (in Chinese). J Forensic Med 22:291–294

Zhang YJ, Zhang HJ, Cui Y, Cui H, Xu QS, Sun S, Sun LP, Lee JB (2007) Population genetics for Y-chromosomal STRs haplotypes of Chinese Korean ethnic group in northeastern China. Forensic Sci Int 173:197–203

Zhang YL (2006) On Zheng He's traveling to the Western world and Chinese immigrant in the Malacca. Acad Forum 3:187–189

Zhao TM, Lee TD (1989) Gm and Km allotypes in 74 Chinese populations: a hypothesis of the origin of the Chinese nation. Hum Genet 83:101–110

Zhu BF, Li XS, Wang ZY, Wu HY, He YF, Zhao J, Liu Y (2005a) Y-STRs haplotypes of Chinese Mongol ethnic group using Y-Plex 12. Forensic Sci Int 153:260–263

Zhu BF, Wang ZY, Yang CH, Li XS, Zhu J, Yang G, Huang P, Liu Y (2005b) Y-chromosomal STR haplotypes in Chinese Uigur ethnic group. Int J Legal Med 119:306–309

Zhu BF, Deng YJ, Zhang FX, Wei WJ, Chen LP, Zhao J, He YF, Tian YF, Xu YC, Yu RJ, Fang JB, Liu Y (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

Zhu BF, Shen CM, Qian GL, Shi RY, Dang YH, Zhu J, Huang P, Xu YC, Zhao QZ, Ma J, Liu Y (2006b) Genetic polymorphisms for 11 Y-STRs haplotypes of Chinese Yi ethnic minority group. Forensic Sci Int 158:229–233

Zhu BF, Liu SZ, Ci D, Huang JF, Wang YC, Chen LP, Zhu J, Xu YC, Zhao QZ, Li SB, Liu Y (2006c) Population genetics for Y-chromosomal STRs haplotypes of Chinese Tibetan ethnic minority group in Tibet. Forensic Sci Int 161:78–83

Zhu BF, Shen CM, Xun X, Yan JW, Deng YJ, Zhu J, Liu Y (2007) Population genetic polymorphisms for 17 Y-chromosomal STRs haplotypes of Chinese Salar ethnic minority group. Leg Med 9:203–209

Zhu BF, Wu YM, Shen CM, Yang TH, Deng YJ, Xun X, Tian YF, Yan JC, Li T (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

Zhu YS, Huo ZH, Yu B, Zhang HB, Wang YJ, Zhao W, Jiao HY, Dang J, Li SB (2007) Genetic polymorphism of twelve Y chromosomal short tandem repeat loci in Chinese Hui ethnic group (in Chinese). Chin J Med Genet 24:594–597