

## ORIGINAL ARTICLE

# Correlation between the linguistic affinity and genetic diversity of Chinese ethnic groups

Hao Sun, Chi Zhou, Xiaoqin Huang, Shuyuan Liu, Keqin Lin, Liang Yu, Kai Huang, Jiayou Chu and Zhaoqing Yang

As the world's most populous nation, China exhibits a population with 56 nationalities. We already know the associations between genetic relationship of these ethnic groups in China and their geographic distributions are closely. However, the correlations between genetic diversity and linguistic affinities have still not been fully revealed in China. To investigate these correlations, 31 populations and 1527 samples were chosen, and the languages of this population covered all of the languages spoken in mainland China (including 8 main linguistic families and 16 subfamilies). The genetic polymorphisms of the populations were investigated using 10 autosomal microsatellites. Five ethnic groups, which included 234 samples, were genotyped in this survey, and the data collected from the other 26 populations were obtained from our previous study. An analysis of molecular variance, principal coordinate analysis, clustering analysis using the STRUCTURE and the Mantel test were used to investigate the correlations between genetic diversity and linguistic affinity. These analyses indicated that most populations who speak the same language demonstrate a similar genetic composition, although a few populations deviated from this linkage between genetics and language. The demographic histories of these populations who deviated from this linkage were investigated. Obvious reasons for why evolutionary processes of genetics and linguistics separated in these populations included geographic isolation, gene replacement, language replacement and intermarriage. Thus, we proposed that the consistency of genetic and linguistic evolution is still present in most populations in China; however, this consistency can be broken by many factors, such as isolation, language replacement or intermarriage.

*Journal of Human Genetics* (2013) 58, 686–693; doi:10.1038/jhg.2013.79; published online 1 August 2013

**Keywords:** genetic diversity; geographic isolation; genetic distances; human evolution; language; population genetics; STR (short tandem repeat)

## INTRODUCTION

The study of human genetic variation can benefit both medical applications and the study of evolution. China has abundant resources for the study of human genetic polymorphisms. China has 1.3 billion people and comprises 56 nationalities, including the Han people (91.6%) and 55 official minority nationalities (8.4%). Although these 55 ethnic groups are known as 'minority nationalities,' they encompass over 100 million people. According to the results of the China population census in 2000, 18 of these ethnic groups have over 1 million members. The largest ethnic group was the Zhuang (16 170 000) in Guangxi Province. Fifteen ethnic groups were of medium size (100 000–1 000 000) and the other 22 ethnic populations were of a smaller size (less than 100 000). The genetic differences among the 55 ethnic groups should not be neglected because of their large population size. The genetic relationships among these groups, and the relationships between the 55 ethnic groups and the Han are still very interesting and important.

Geographic distance and linguistic boundaries both limit gene flow and maintain the genetic differences between the populations. By investigating the polymorphism of microsatellites in autosomes,<sup>1</sup> human leukocyte antigen<sup>2</sup> and mitochondrial DNA,<sup>3</sup> many studies have already shown a significant correlation between genetics and geographic distances in Chinese ethnic groups.<sup>1,4–6</sup> Thus, geographic distance has been considered a main factor that limits gene flow in China. In 2005, after comparing the genetic relationships of 51 populations worldwide, Manica *et al.*<sup>7</sup> suggested that 'pair-wise geographic distances across landmasses constitute a far better predictor than ethnicity'. If this is true in China, then does it mean that we can infer that genetic relationships between ethnic groups are based on their geographic distances and that the ethnic or linguistic backgrounds may be ignored? The correlations between genetic diversity and linguistic affinities are still not fully understood in China. Thus, we aimed to investigate these correlations and how geographic distance affects these relationships.

In our previous study,<sup>8</sup> we found that some aggregation patterns of populations in China corresponded with the populations' linguistic affiliations. However, these patterns were unclear, and three important language families in China were not included. These languages included the Korean, Hmong-Mien and Altaic-Tungusic language families. Thus, five ethnic groups who speak Hmong-Mien, Altaic-Tungusic and Korean were selected, and thus, all of the existing languages in mainland China were included in our survey. The five ethnic groups were the She (Hmong-Mien, Ho Nte), Yao (Hmong-Mien, Hmongic), Ewenki (Altaic, Tungusic), Manchu (Altaic, Tungusic) and Korean (Korean, Language isolate), which produced a total of 234 samples. We used the same 10 microsatellites as in our previous report<sup>8</sup> to allow amalgamation of the data.

Microsatellite markers in autosomes have been used to calculate genetic distances and have been proven as ideal markers to study closely related populations.<sup>9,10</sup> Microsatellite markers can generally provide a reliable genetic relationship. Compared with single-nucleotide polymorphisms on the Y chromosome and mitochondrial DNA, autosomal microsatellites carry the information from both parents. The ease of genotyping using microsatellite alleles have made them a good tool to use in the study of population genetics.

## MATERIALS AND METHODS

### Sampled populations and DNA preparation

In this study, 234 people were recruited from five ethnic groups (She, Yao, Ewenki, Manchu and Korean in China), who speak Hmong-Mien, Altaic and Korean, for further genotyping. The genotyping data obtained from the other 26 populations were collected from our previous study.<sup>8</sup> The locations and language classifications of these ethnic groups are indicated in Figure 1 and Table 1. As shown in Figure 1, these populations were generally divided into five groups according to their geographic location; Area A was northwestern China, Area B was northeastern China; Area C was central China, which was a region mainly situated around the Yellow River Basin, where Han people

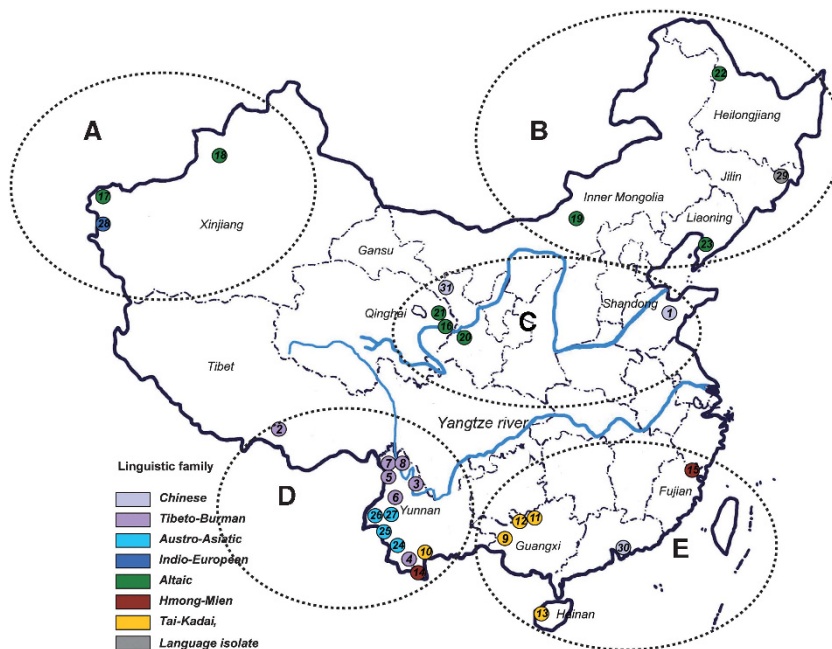
developed and grew in the history of China. Area D was southwest of China and Area E was southeast of China. In general geographic terms, China was divided into the south and the north by Yangtze River, and Area A, B and C were regions that were all located north of the Yangtze River and were considered as northern China. Area D and E were southern China. When compared with Area E, Area D had a more complex terrain. Salween River, Mekong River, Jingsha River, Hengduan Mountains and Yunlin Mountains all locate in Area D. High mountains and a complex river system blocked the gene flow.

Samples were collected via a coordinated effort of several institutes participating in the Chinese Human Genome Diversity Project.<sup>11</sup> All of the blood contributors consisted of local people whose parents and grandparents belonged to the same ethnic group.

All of the DNA samples were obtained from immortalized cell lines. A written informed consent had been signed for the establishment of the cell lines as well as for the use in subsequent studies. The Ethics Committee at the Chinese Academy of Medical Sciences and Peking Union Medical College approved this project. The DNA of the five ethnic groups and of the four other ethnic groups (Han (Shandong), Tibetan, Wa (Ximeng) and Wa (Canyuan)) with incomplete data in our previous study<sup>8</sup> was extracted using the DNA Miniprep Kit (Axygen, Tewksbury, MA, USA).

### Genotyping of 10 microsatellites

The same 10 microsatellites used in our previous study<sup>8</sup> were chosen for genotyping in this study. These microsatellites were obtained from the ABI Prism Linkage Mapping Set. They were localized on to chromosome 3 and the mean distance of the adjacent markers was 23.5 cM. The large distances between the adjacent markers were good for minimizing the effect of linkage disequilibrium on the genetic distance calculation. All of the 10 microsatellite markers were CA repeats and highly informative. All of the 10 markers were genotyped in the five ethnic groups in this study. Microsatellite D3S1580 was genotyped in Wa (Canyuan) and Wa (Ximeng), and D3S1304 was genotyped in Han (Shandong) and Tibetan because they were missing in the study conducted by Lin *et al.* 2010. Twenty samples which have been genotyped in Lin's paper were analyzed again in this survey, in order to make sure that the data were not affected by potential batch effect.



**Figure 1** Geographical location of the 31 sampled populations. The numbers in the figure represent the number of populations in Table 1. These populations were generally divided into five groups according to their geographical location. Area A represents northwest China. Area B represents northeast China. Area C represents central China, and this region was mainly around the Yellow River Basin, where the Han people developed and grew according to the Chinese history. Areas D and E represented southwest and southeast China. Compared with Area E, Area D demonstrated more complex terrain.

**Table 1** Demographic and linguistic information of the 31 sampled populations

No.	Population	Size	Location	Lat(N)	Long(E)	Language classification
1	HanSD	45	Zouping, Shandong	36.86	117.74	Chinese
2	Tibetan	46	Ihasa, Tibetan	29.66	91.13	Tibeto-Burman, Himalayish, Tibeto-Kanauri, Tibetic, Tibetan
3	Yi	62	Ninglang, Yunnan	27.28	100.75	Tibeto-Burman, Burmic, Ngwi, Northern
4	Hani	55	Jinghong, Yunnan	22.01	100.79	Tibeto-Burman, Burmic, Ngwi, Southern
5	Lisu	56	Fugong, Yunnan	25.84	98.85	Tibeto-Burman, Lolo-Burmese, Loloish, Northern, Lisu
6	Pumi	69	Lanping, Yunnan	26.45	99.42	Tibeto-Burman, Tangut-Qiang, Qiangic
7	Dulong	55	Gongshan, Yunnan	27.74	98.67	Tibeto-Burman, Nungish
8	Nu	52	Gongshan, Yunnan	27.74	98.67	Tibeto-Burman, Nungish
9	Zhuang	56	Baise, Guangxi	23.9	106.62	Tai-Kadai, Kam-Tai, Be-Tai, Tai-Sek, Tai, Northern
10	Dai	60	Jinghong, Yunnan	22.01	100.79	Tai-Kadai, Kam-Tai, Be-Tai, Tai-Sek, Tai, Central
11	Mulam	52	Luocheng, Guangxi	24.91	108.84	Tai-Kadai, Kam-Tai, Kam-Sui
12	Maonan	52	Hechi, Guangxi	24.83	108.26	Tai-Kadai, Kam-Tai, Kam-Sui
13	Li	53	Baisha, Hainan	19.23	109.45	Tai-Kadai, Hlai
14	Yao <sup>a</sup>	51	Mengla, Yunnan	21.48	101.57	Hmong-Mien, Hmongic, Bunu
15	She <sup>a</sup>	45	Fuan, Fujian	27.09	119.65	Hmong-Mien, Ho Nte
16	Salar	53	Xunhua, Qinghai	35.87	102.43	Altaic, Turkic, Southern
17	Kirgiz	45	Wuqia, Xinjiang	39.71	75.25	Altaic, Turkic, Western, Aralo-Caspian
18	Uyghur	42	Yili, Xinjiang	43.92	81.32	Altaic, Turkic, Eastern
19	Mongolian	50	Damaoqi, Inner Mongolia	41.70	110.43	Altaic, Mongolic, Eastern, Oirat-Khalkha, Khalkha-Buriat,
20	Dongxiang	45	Dongxiang, Gansu	35.66	103.39	Altaic, Mongolic, Eastern, Mongour
21	Tu	37	Huzhu, Qinghai	36.84	101.95	Altaic, Mongolic, Eastern, Mongour
22	Ewenki <sup>a</sup>	38	Nehe, Heilongjiang	48.48	124.88	Altaic, Tungusic, Northern, Evenki
23	Manchu <sup>a</sup>	50	Xiuyan, Liaoning	40.28	123.29	Altaic, Tungusic, Southern, Southwest
24	WaCangyuan	50	Cangyuan, Yunnan	23.15	99.25	Austro-Asiatic, Mon-Khmer, Northern Mon-Khmer, Palaungic
25	WaXimeng	49	Ximeng, Yunnan	22.64	99.6	Austro-Asiatic, Mon-Khmer, Northern Mon-Khmer, Palaungic
26	Bulang	39	Luxi, Yunnan	23.44	98.59	Austro-Asiatic, Mon-Khmer, Northern Mon-Khmer, Palaungic
27	Deang	52	Luxi, Yunnan	23.44	98.59	Austro-Asiatic, Mon-Khmer, Northern Mon-Khmer, Palaungic
28	Tarjike	40	Tashikuergan, Xinjiang	37.77	75.23	Indo-European, Indo-Iranian, Iranian, Eastern, Southeastern, Pamir
29	Korean <sup>a</sup>	50	Yanbian, Jilin	42.91	129.52	Korean (Language isolate)
30	HanGD	39	Guangning, Guangdong	23.63	112.44	Chinese
31	HanGS	39	Weiwu, Gansu	37.92	102.63	Chinese
	Total	1527				

Lat and Long represent latitude (north) and longitude (east), respectively.

<sup>a</sup>Population samples from this study. The other 26 population sample groups were published previously in Lin *et al.*<sup>8</sup>

PCR reactions contained 10–20 ng of genomic DNA, 0.2  $\mu\text{mol l}^{-1}$  of primer mix, 200  $\mu\text{mol l}^{-1}$  of each deoxynucleotide triphosphate (dNTP), 1  $\times$  PCR buffer and 0.1U Transtart DNA polymerase (Transgen, Peking, China) in a total volume of 5  $\mu\text{l}$ . Amplification reactions were performed in a Perkin Elmer GeneAmp PCR System 9600 thermal cycler (Applied Biosystems, Foster City, CA, USA), and the cycling conditions were as follows: initial denaturation at 94 °C for 5 min, followed by 10 cycles of 94 °C for 30 s, 55 °C for 30 s and 72 °C for 30 s, 25 cycles of 89 °C for 30 s, 55 °C for 30 s and 72 °C for 30 s and a final extension at 72 °C for 10 min. Detection of the amplified products was accomplished using the ABI3130 genetic analyzer (Applied Biosystems), and the microsatellite alleles were represented by their product size. The data files were generated using ABI PRISM GeneScan Analysis (Applied Biosystems) and were analyzed using the GeneMarker software (SoftGenetics, State College, PA, USA).

### Data analysis

Arlequin version 3.11<sup>12</sup> was used to calculate the expected heterozygosity ( $H_E$ ),<sup>13</sup> linkage disequilibrium of adjacent loci and Hardy–Weinberg Equilibrium of the 10 markers in the five ethnic groups. The Fixation index  $F$  ( $F_{ST}$ )<sup>14</sup> between paired populations was also calculated using Arlequin; however, their significances were tested by 1000 permutations.

To reveal the patterns of the genetic relationship contained in the  $F_{ST}$  matrix, the principal coordinate analysis (PCoA) was performed using GenALEx.<sup>15</sup> PCoA is a process in which the major axes of variation were located within a multidimensional data set. For the distinct groups, the first two or three axes

will typically reveal most of the separation between the groups. A three-dimensional scatter chart was drawn on the basis of the PCoA analysis, which was obtained using SPSS Statistics 17.0 (IBM, Armonk, NY, USA).

To further quantify the effect of language and geography on the genetic relationship of these populations, analysis of molecular variance (AMOVA) at different levels of the hierarchical subdivision was performed using Arlequin.<sup>16</sup> Thirty-one populations were grouped according to their linguistic affiliation or geographic location as indicated in Figure 1. Differences in their genetic significances were tested by 10 000 permutations.

The Mantel test can be used to test the significance of the correlation between two or more matrices.<sup>17,18</sup> To evaluate the correlation among the genetic, linguistic and geographic matrices, the correlation and partial correlation were tested using the Mantel test in the Arlequin program. The Slatkin's linearized  $F_{ST}$ <sup>19</sup> between paired populations was considered a genetic distances matrix. However, the geographic distances matrix was calculated from the latitude/longitude coordinates using the modified Haversine formula, which was developed by Sinnott<sup>20</sup> in the GenALEx program, and was transformed into  $\log(1 + X)$ . The linguistic distances were determined according to the 'least controversial phylogeny', which was proposed by Sagart *et al.*<sup>21</sup> In this study, the Hmong-Mien was considered an independent branch similar to the Tai-Kadai, according to the language classification system defined in Gordon's and Grimes book.<sup>22</sup> Thus, similar to the Tai-Kadai, the age of the most recent common ancestor of the Hmong-Mien was 50 000. The permutations of 1000 were performed during the Mantel test.

To demonstrate the characteristics of population structure, we used a model-based clustering method in the STRUCTURE V2.2 program<sup>23,24</sup> to estimate the reasonable partitions of these populations. The parameters of the STRUCTURE were established according to the suggestion of Falush *et al.*<sup>25</sup> The degree of the admixture  $\alpha$  was inferred from the data, the admixture model was chosen and the correlations of allele frequencies between populations were taken into account. The location of the samples (LOCPRIOR model<sup>26</sup>) was used to detect weak population structures. Population structures were inferred by setting the value of the clusters ( $K$ ) from 2 to 7. Five runs were performed for every  $K$ -value with an Markov chain Monte Carlo chain burn-in length of 80000 iterations followed by 80000 iterations.  $K$  was identified using the value of the average logarithmic probability across runs returned by the STRUCTURE 2.3. This  $K$  is an estimate of the optimum  $K$  for accurate representation of the ancestry. Outputs from the STRUCTURE were graphically modified by DISTRUCT.<sup>27</sup>

## RESULTS

### Standard diversity indices

The allele frequencies,  $H_E$  and Hardy–Weinberg Equilibrium test results of the 10 microsatellites in the five ethnic groups are shown in Supplementary Table S1. The mean  $H_E$  of all the 10 markers was greater than 0.6 in the five ethnic groups. The lowest  $H_E$  was 0.625 (D3S1266 of She) and the highest  $H_E$  was 0.901 (D3S1292 of Manchu). Thus, all the 10 markers were highly diversified in the five ethnic groups. After the Bonferroni correction for multiple tests, all of the loci were consistent with Hardy–Weinberg Equilibrium in the five ethnic groups at a significance level of  $\alpha = 0.001$ . The missing data rates of all of the polymorphic loci were less than 5%. The results of 10 STRs in 20 resequencing samples were same with our earlier work,<sup>8</sup> so the data from different sources were not affected by potential batch effect.

To ascertain whether the linkage disequilibrium affects our analysis, linkage disequilibrium of the adjacent loci was tested in the five populations. Linkage disequilibrium beyond the adjacent loci was not tested because of the high recombination fractions between loci (for example, an average of 23.5 cM for next-to-adjacent loci), which would be unlikely to produce linkage disequilibrium. After the Bonferroni correction for multiple tests, all of the loci in the five ethnic groups were not linked at the  $\alpha = 0.001$  significance level.

### Fixation index $F$ and PCoA

The results of the  $F_{ST}$  between paired populations are shown in the lower triangle in Supplementary Table S2. The significance of  $F_{ST}$  was tested using 1000 permutations. The  $P$ -values are shown in the upper triangle of Supplementary Table S2. Each paired  $F_{ST}$  was less than 0.1. The biggest  $F_{ST}$  was observed between the Dulong and Tarjike (0.0924). In addition, the smallest  $F_{ST}$  was 0.0005, which was between the Salar and Dongxiang. After a Bonferroni correction for multiple tests, most of the paired populations demonstrated a significant difference ( $P < 0.0001$ ). When comparing the  $F_{ST}$  with another type of marker such as the two allelic single-nucleotide polymorphism,<sup>28</sup> the  $F_{ST}$  in our survey was small. This may be due to hypervariable markers such as the microsatellites used, which have a high  $H_E$ . Thus, an  $F_{ST}$  based on these microsatellites will be small.<sup>29</sup> However, whether a significant genetic differentiation ( $F_{ST} > 0$ ) may be detected is of importance. In our study, most of the  $F_{ST}$  among the 31 populations were significantly greater than 0. Thus, there were genetic differences among most of the ethnic groups in China.

We also noticed that three new populations (Yao, She and Korean), Pumi, Dulong, Nu, Li, Mongolian, Wa\_Cangyuan and Tarjike are significantly different from all others by  $F_{ST}$  analyses Supplementary

Table S2. In other word, these populations appear unique by  $F_{ST}$  analyses.

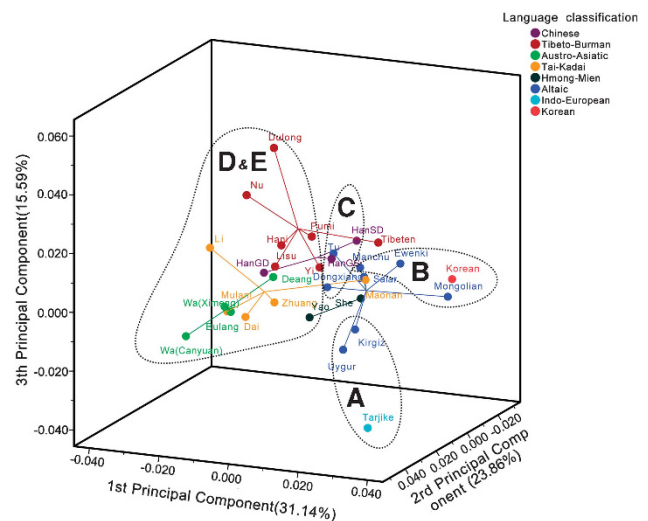
A three-dimensional scatter chart was drawn on the basis of the PCoA analysis to better describe the genetic relationships of these populations (Figure 2). The first principal component was 31.14%, the second principal component was 23.86% and the third principal component was 15.59%. Thus, the three-dimensional scatter chart can represent  $\sim 71\%$  of the total genetic variations. To investigate the relationships among genetics, language and geography, the populations shown in Figure 2 were colored according to their linguistic affiliation. Areas A–E, which were marked in Figure 2, demonstrated the geographic locations of these populations. These areas were previously described and shown in Figure 1.

The ethnic groups who spoke the same language were generally clustered together. In Figure 2 the genetic distance of the populations can be divided into three main parts in the PCoA plot. The Northern Han and populations of Tibeto-Burman were distributed in the upper right of the plot. The ethnic groups of the Austro-Asiatic and Tai-Kadai were clustered in the lower left and the ethnic groups of the Altaic, Korean and Tarjike of Indo-Europe were clustered in the lower right. The Hmong-Mien appeared to have a closer genetic distance with the Altaic. Thus, the genetic variations of the ethnic groups in China could be partially grouped according to their linguistic affiliation.

The relationship between genetic variations and geographic locations are shown in Figure 2. The populations in northern China were grouped according to their geographic locations (Area A, B and C). In contrast, people who lived in Southern China (Area D and E) did not show significant clustering. Thus, the genetic distance among the populations in Southern China could not be distinguished on the basis of their geographic locations.

### The AMOVA and Mantel test

The results of the AMOVA at different levels of the hierarchical subdivision are shown in Table 2. The 31 populations were grouped according to their linguistic affiliation or geographic locations.



**Figure 2** PCoA with  $F_{ST}$  values. The percentages of variance accounted for by the three components are indicated in the labels. The populations were colored according to their linguistic affiliations to obtain a better visual comparison. Their geographical locations (shown in Figure 1) were marked under the picture.

**Table 2 Results of AMOVA**

Populations used to analyze	Grouped by	Within populations		Among populations within groups		Among groups	
		Variance (%)	$F_{ST}$	Variance (%)	$F_{SC}$	Variance (%)	$F_{CT}$
All populations	Language	97.16	0.028	1.90	0.019	0.94	0.009
	Geography	97.11	0.029	2.17	0.022	0.72	0.007
North populations	Language	97.34	0.027	1.50	0.015	1.17	0.012
	Geography	97.49	0.025	1.25	0.012	1.26	0.013
South populations	Language	97.27	0.027	2.13	0.021	0.60	0.006
	Geography	97.40	0.026	2.60	0.026	0.00	—

Abbreviation: AMOVA, analysis of molecular variance.

All  $F$  statistics are significant and their  $P$ -values are less than 0.0001. (The  $\alpha$  was set as 0.003 after the Bonferroni correction).

**Table 3 Correlation coefficients ( $r$ ) among genetic (*Gen*), geographic (*Geo*) and linguistic (*Lan*) matrix calculated by the Mantel test**

	$r_{Gen*Geo}$ (P)	$r_{Gen*Lan}$ (P)	$r_{Geo*Lan}$ (P)	$r_{(Gen*Geo)Lan}$ (P)	$r_{(Gen*Lan)Geo}$ (P)	$V_{Geo}$ (%)	$V_{Lan}$ (%)
All populations in the survey	0.248 (0.008)	0.192 (<0.001)	0.295 (<0.001)	0.204 (0.020)	0.129 (0.030)	5.2	2.5
Populations of North China	0.507 (<0.001)	0.361 (0.070)	0.060 (0.291)	0.521 (<0.001)	0.383 (0.050)	25.7	—
Population of South China	0.150 (0.108)	0.059 (0.330)	0.181 (0.034)	0.142 (0.136)	0.032 (0.415)	—	—

$P$ -values less than 0.003 (The  $\alpha$  was set as 0.003 after the Bonferroni correction) were shown in bold.  $r_{(Gen*Geo)Lan}$  is a partial correlation coefficients, which means the correlation between *Gen* and *Geo* when the influence of *Lan* was considered.  $r_{(Gen*Geo)Lan}$  means the correlation between *Gen* and *Lan* when the influence of *Geo* was considered.  $V_{Geo}$  is the percent of genetic variation which is determined by geographical distances (*Geo*).  $V_{Lan}$  is the percent of genetic variation determined by *Lan*.

Because previous studies have suggested that populations in northern China and in southern China demonstrate different genetic backgrounds<sup>1</sup> and that they may migrate into East Asia from two different roads similar to a 'pincer',<sup>2,30</sup> the AMOVA test was also performed on the northern and southern populations of China.

In these analyses, all of the  $F$ -values were significantly larger than 0. However, the  $F$ -value among the groups ( $F_{CT}$ ) was always smaller than the  $F$ -value among the populations within groups ( $F_{SC}$ ) for most of the comparisons, with the exception of the analysis of the northern populations that were grouped according to their geographical locations. This indicated that the genetic variances among the groups were smaller than the variance among the populations within groups. Thus, the linguistic affiliation of populations does not link well with the genetic variances of these populations, and the correlation between genetics and geography may only be observed in the northern populations of China.

Mantel tests were also performed to reveal the relationships among genetics, linguistics and geographic distances. Correlation coefficients and significance Mantel tests among these three matrices were calculated, and the results are shown in Table 3. Similar results were obtained using an AMOVA.

For all of the 31 populations in China, the correlation coefficient between the genetic distances and geographic distances was not significant ( $r_{Gen*Geo} = 0.248$ ,  $P = 0.008$ ). However, after Bonferroni corrections, the correlation coefficient between the genetic distances and language was significant ( $r_{Gen*Lan} = 0.192$ ,  $P < 0.001$ ). After analyses of the partial correlation, which considered the effect of geographical distances, these significant differences disappeared. Thus, the correlation between genetics and language or geography was not very strong for all of the populations.

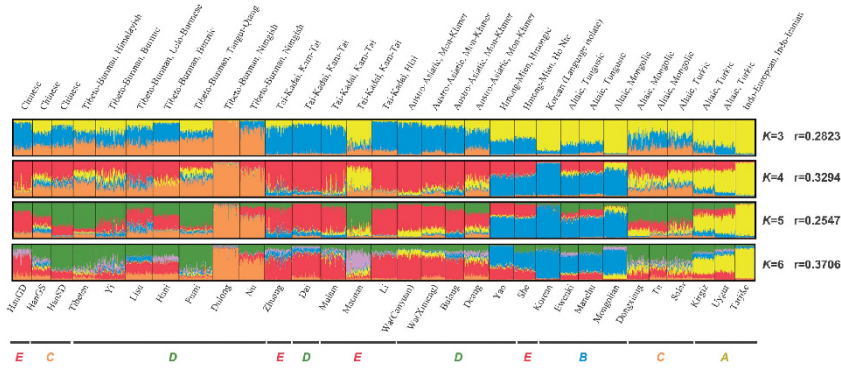
For populations of northern China, the correlation was only significant between the genetic distances and geographic distances ( $r_{Gen*Geo} = 0.507$ ,  $P < 0.001$ ). This significance was still present after analyses of the partial correlation. For populations of southern China,

the correlation was not significant between both the genetic distances and geographic distances and between the genetic distances and linguistic affiliation.

#### Cluster analysis using the STRUCTURE

The population structure was inferred using the STRUCTURE program. The population structures of a representative run at  $K = 3-6$  are shown in Figure 3. The output posterior probabilities ( $\ln Pr(X|K)$ ) was highest at  $K = 5$  (Supplementary Figure S1). According to the instructions in the STRUCTURE manual,  $K = 5$  was the most appropriate configuration. In Figure 3, each individual was represented by a single vertical line that was broken into different  $K$ -colored segments. The lengths of these different colored segments were represented the percentage of genetic components inferred by STRUCTURE. When the LOCPRIOR model was used, the  $r$ -value was reported. An  $r$ -value less than 1 indicated that the sampling location in the model was effective.<sup>26</sup> The 31 populations were arrayed by linguistic affiliation and their geographic locations were marked under the picture.

The proportions of genetic components were different in the same language family on the basis of the results of the STRUCTURE. When  $K = 5$ , there were five different genetic components, which were represented by five different colors, in these populations. The red-colored ancestral component inferred using the STRUCTURE was predominant in the south Han (HanGD) people and in populations that spoke Tai-Kadai, Austro-Asiatic. The green-colored ancestral component was predominant in the north Han (HanGS and HanSD) and in populations that spoke Tibeto-Burman. The blue-colored ancestral component was predominant in populations that spoke Hmong-Mien and Korean. The yellow-colored ancestral component was predominant in populations that spoke Indo-European and in Altaic speakers who lived in the northwestern part of China (Area A in Figure 1). The orange-colored ancestral component was predominant in the two Tibeto-Burman-speaking populations.



**Figure 3** Clustering analysis by the STRUCTURE assuming  $K=3, 4, 5$  and  $6$ . The populations were ordered according to their linguistic affiliations. The linguistic affiliations and population names were labeled above and below the plot, respectively. The name of the ethnic groups and their geographical locations (shown in Figure 1) were marked under the picture. The  $r$ -value to the right of each Distruct plot is the structure parameter for estimation of the informativeness of sampling location.

The geographical locations of these populations were marked in the STRUCTURE results in Figure 3. Populations in the same geographical locations (Area A, B or C) demonstrated a similar genetic makeup in northern China. In contrast, in southern China, populations residing in the same locations (Area D or E) had a distinctly different genetic makeup.

### DISCUSSION

We aimed to investigate whether the ethnic or linguistic backgrounds of these populations were useful to infer their genetic backgrounds in China. On the basis of the AMOVA results and Mantel test for all 31 populations, there were genetic differences among the different language speakers. However, the correlation between genetic diversity and linguistic affinity was not very significant. Although these 31 populations were divided into two groups, northern and southern populations for the analysis, the correlation was still difficult to obtain.

Both genetic material and language can be passed down by the parents of offspring. The origins of populations or demographic changes can often affect both the evolutionary process of genetic and language. Thus, genetics and language usually share a similar evolutionary process. However, their correlation may be decreased or hidden by other phenomena, such as language replacement or gene replacement.<sup>31</sup> To determine what factors decreased the correlation between genetics and language in China, a more detailed analysis was performed. The proportion of genetic components of all 31 populations was inferred using the STRUCTURE software. According to the results of the STRUCTURE analysis, we will discuss the factors that decreased the correlations between genetics and language in different language families.

#### Chinese and Tibeto-Burman speakers

In linguistics, the Chinese and Tibeto-Burman have a very close relationship. Some linguists have suggested that these two language families may be combined to form a super language family: the Sino-Tibetan. Thus, populations speaking Chinese and Tibeto-Burman are discussed together. There were 10 populations in this group, including three Chinese-speaking Han populations and seven other Tibeto-Burman-speaking populations. Most of these populations share a similar genetic makeup; however, the HanGD, Nu and Dulong showed a different genetic makeup.

Three Chinese-speaking Han populations were selected to participate in this survey, including two typical northern Han (HanGS and

HanSD) and a typical southern Han (HanGD). We found that the northern Han demonstrated a similar genetic structure with people speaking Tibeto-Burman. Furthermore, the genetic structure of the southern Han was similar with people speaking Tai-Kadai. In Chinese history, the Han people had continuously expanded southward with a demic diffusion model.<sup>3</sup> Thus, the southern and northern Hans have a different genetic makeup, and northern immigrants and southern natives were combined to form the current southern Han population. The HanGD have a different genetic makeup compared with the northern Han and a similar genetic makeup compared with the southern Tai-Kadai speakers.

People who speak Tibeto-Burman share a similar genetic structure except for the Dulong and Nu on the basis of the STRUCTURE results. The ancestral component represented in green was predominant in most Tibeto-Burman people; however, the orange ancestral component was predominant in the Nu and Dulong people. The Dulong and Nu people live in close proximity with other people who speak Tibeto-Burman, and there is no evidence that they have a different origin from people who speak Tibeto-Burman. After investigating the history of the Dulong and Nu, we found that these people were driven into the Valley of the upper reaches of the Salween River (Nu River in China) by powerful tribes. The area where the Dulong and Nu inhabited has a complex topography. Until 1999, there was only a simple road that led to the residential area of Dulong people, and in the winter this road was blocked by heavy snow for at least 6 months. High mountains and a complex river system blocked the gene flow. Thus, we believe that the genetic differences between the Dulong or Nu and other people who speak Tibeto-Burman were mainly caused by geographical isolation, particularly, topographical isolation.

#### Tai-Kadai and Austro-Asiatic speakers

The Tai-Kadai and Austro-Asiatic are both important language families in southern China. They share a similar genetic makeup and thus, we will discuss these two groups together. The red ancestral component inferred using the STRUCTURE was predominant in people who speak Tai-Kadai and Austro-Asiatic. This may imply that people who speak Tai-Kadai and Austro-Asiatic shared a common ancestor. Other reports on the basis of the variations on the Y chromosome and mitochondria have also suggested that these two groups have a genetic affinity and suggested that these two linguistic groups were genetically distinct, but only by a marginal fraction (1–2%) of the total genetic variation.<sup>32,33</sup>

However, there is an exception; the Maonan people who speak Tai-Kadai show a different genetic makeup compared with other southern populations. It is difficult to explain this difference using the origin or geographical isolation. However, when a deep investigation was performed, we found that the Maonan people had intended to marry within their ethnic group. There were ~70% Maonan people who had the same family name 'Tan.' Thus, we suggested that the intermarriages between Maonan people might have resulted in their distinct genetic structure. Inbreeding in this ethnic group may have maintained their unique genetic makeup.

### Hmong-Mien speakers

The genetic structure of people who speak Hmong-Mien (She and Yao) was very interesting. These two populations who speak Hmong-Mien are both located in southern China, although their genetic structures are very similar with Altaic speakers, who mostly live in northern China. The blue ancestral component inferred using the STRUCTURE was predominant in this population. The correlation between genetics and geography was broken for speakers of Hmong-Mien. Similar results were reported using different genetic markers. The principal component analysis on the basis of the haplotype frequencies of genetic markers on the Y chromosome indicated that the She and Yao have a close genetic distance with the speakers of Altaic.<sup>34</sup> A multidimensional scaling analysis of  $F_{ST}$  on the basis of single-nucleotide polymorphisms on the Y chromosome was performed by Karafet *et al.*,<sup>28</sup> which also showed similar results. Wen *et al.*<sup>35</sup> proposed that a higher frequency of north-dominating lineages was observed in the Hmong people using the mitochondrial DNA lineage analysis.<sup>35</sup>

Some historical events may explain this phenomenon. In ancient times, the ethnic groups of the Hmong-Mien originated from the 'Man' people who lived in the Jiangnan Plains (Central China). According to a famous Chinese legend, some ancient 'Man' people (the *San-Miao* tribe), who were thought to be the ancestor of the Hmong-Mien populations, had expanded northward to the Yellow River Valley. On their northward migration, the *San-Miao* tribe led by *Chiyou* had battled against the north *Yan-Huang* tribe (the primary Sino-Tibetan ancestors), who were led by *Huangdi*. *Chiyou* and his tribe were defeated and were pushed back to the south.<sup>36</sup> This may explain why the Hmong-Mien speakers are located in southern China but have a similar genetic makeup with the northern population.

### Altaic speakers and other northern populations

We analyzed eight Altaic-speaking populations and two other populations (Korean and Tarjike) in this survey. These northern populations were divided into three groups on the basis of their genetic structure. The two Altaic-speaking populations (Kirgiz and Uygur) and the Tarjike people who speak Indo-European have a similar genetic makeup. The yellow ancestral component inferred using the STRUCTURE was predominant in this population. Three Altaic-speaking populations (Ewenki, Manchu and Mongolian) showed a similar genetic makeup with the Korean people. The blue ancestral component was predominant in this population. The other three Altaic-speaking populations (Dongxiang, Tu and Salar) demonstrated a similar genetic makeup with the northern Han people. Interestingly, these three genetic different groups lived in three different geographical locations in northern China (Area A, B and C in Figure 1). The correlation between languages and genetic was broken in northern China.

The Dongxiang, Tu and Salar shared a similar genetic makeup with the northern Han people who lived in central China (Area C in

Figure 1). We proposed that language replacement<sup>31</sup> might be observed in these people. According to Chinese history, central China (Yellow River Basin) was fertile. The Han people lived on in a village here, but the northern Altaic-speaking nomads had repeatedly invaded the Yellow River Basin over the course of Chinese history. Superior military power made the Altaic to become the dominant language in this area. People who spoke Chinese were beginning to use Altaic, and thus, language replacement occurred. There are history records that support this proposal. Some ethnologists also believe that some populations in the Yellow River Basin, such as the Dongxiang, Salar and Tu, had originated from a mixed population of Han, Tibetan and a small number of Mongolian people.<sup>37</sup> Thus, we can understand why people who speak Altaic have a similar genetic structure with the Han and Tibetan people who live in the Yellow River Basin. Here the correlation between genetics and language was broken by language replacement.

The Kirgiz, Uygur and Tarjike live in northwestern China (Area A in Figure 1). The Altaic-speaking Kirgiz and Uygur have a different genetic makeup with other Altaic speakers but have a similar genetic makeup with the Indo-European-speaking Tarjike. We proposed that the isolation of geographical distance might have caused this phenomenon. Not only was the geographical distance between the Altaic speakers living in Area A and other Altaic speakers (mainly living in northeastern China, Area B in Figure 1) very far, but there were also some vast deserts, such as the Taklimakan desert, in the middle of these two regions. Thus, the gene flow, which easily occurs between the same language speakers, was cutoff. The gene flow that originated from the Indo-European speakers was the main influencing factors of the Kirgiz and Uygur gene pools. The correlation between genetics and language was broken by this 'gene replacement'<sup>31</sup> or isolation in geographical distance.

### CONCLUSION

In this study, the correlations of 31 populations' genetics, language and geography were investigated. These 31 populations represented approximately half of the identified ethnic groups in China, and their languages covered all of the languages spoken in mainland China. Although the number of samples obtained from a few populations was less than 40, the nature of these populations was still well presented.

These results showed that if these 31 populations were studied as an entire picture, the correlation between genetics and language was weak. Although these populations were divided into two groups, the southern groups and northern groups, the correlation was still difficult to observe. However, when a structure analysis was performed for every population, we found that just a few populations showed different genetic compositions in populations who spoke the same language. Demographic and geographic investigations suggested that geographical isolation, gene replacement, language replacement and intermarriage can alter the genetic compositions of these populations. Thus, in China, the correlation between genetics and language can be broken, but in most populations who speak the same language, the people still share similar genetic compositions as assessed using microsatellite analyses.

In our results, we also noticed that although some populations appear to share patterns of ancestry from the STRUCTURE analysis, they are different by  $F_{ST}$  analyses. These discrepancies can be found among Yao, She, Korean and Mongolian. Similar discrepancies also can be found between Dulong and Nu or between Li and Wa\_Canyuan. To explain this discrepancy, we proposed that gene flow between these contemporary populations has been restricted long

enough for significant differences to accrue, but the genetic signal of their shared ancestry is still discernible.

The correlation between genetics and geography was also investigated in this study and was only found in northern China. This correlation was also decreased or hidden by the previously described phenomena. An interesting sample was the Hmong-Mien-speaking people. In our study, the two ethnic groups (the She and Yao) who speak Hmong-Mien have a large geographical distance between each other as well as a large geographical distance with the Altaic-speaking people, such as the Mongolian and Ewenki people. However, these populations have a similar genetic makeup. Therefore, when the genetic background information is needed for a case-controlled study or other related studies, the effect of both geography and language should be considered. However, a preliminary genetic investigation will be a better option.

In this study, only 31 populations participated. Thus, to reveal a more complete picture of the genetic, linguistic and geographic relationship, a comprehensive study involving all Chinese ethnic populations would be necessary.

#### ACKNOWLEDGEMENTS

This work was supported by Youth Foundation of Chinese Academy of Medical Sciences and Peking Union Medical College (No. 3332013085), in part by Yunnan Science and Technology Program (NO. 2008ZC068M) and Chinese National High Technology Research and Development Program (No. 2012AA021802).

- Chu, J. Y., Huang, W., Kuang, S. Q., Wang, J. M., Xu, J. J., Chu, Z. T. *et al*. Genetic relationship of populations in China. *Proc. Natl Acad. Sci. USA* **95**, 11763–11768 (1998).
- Di, D. & Sanchez-Mazas, A. Challenging views on the peopling history of East Asia: the story according to HLA markers. *Am. J. Phys. Anthropol.* **145**, 81–96 (2011).
- Wen, B., Li, H., Lu, D., Song, X., Zhang, F., He, Y. *et al*. Genetic evidence supports demic diffusion of Han culture. *Nature* **431**, 302–305 (2004).
- Abdulla, M. A., Ahmed, I., Assawamakin, A., Bhak, J., Brahmachari, S. K., Calacal, G. C. *et al*. Mapping human genetic diversity in Asia. *Science* **326**, 1541–1545 (2009).
- Sun, H., Zhou, C., Huang, X. Q., Lin, K. Q., Shi, L., Yu, L. *et al*. Autosomal STRs provide genetic evidence for the hypothesis that Tai People originate from Southern China. *PLoS One* **8**, e60822 (2013).
- Yao, Y. G., Nie, L., Harpending, H., Fu, Y. X., Yuan, Z. G. & Zhang, Y. P. Genetic relationship of Chinese ethnic populations revealed by mtDNA sequence diversity. *Am. J. Phys. Anthropol.* **118**, 63–76 (2002).
- Manica, A., Prugnolle, F. & Balloux, F. Geography is a better determinant of human genetic differentiation than ethnicity. *Hum. Genet.* **118**, 366–371 (2005).
- Lin, H., Fan, H., Zhang, F., Huang, X., Lin, K., Shi, L. *et al*. Genetic relationships of ethnic minorities in Southwest China revealed by microsatellite markers. *PLoS One* **5**, e9895 (2010).
- Nei, M. & Takezaki, N. The root of the phylogenetic tree of human populations. *Mol. Biol. Evol.* **13**, 170–177 (1996).
- Wright, S. *Evolution and the Genetics of Populations; a Treatise* (University of Chicago Press, Chicago, Illinois, USA, 1968).
- Cavalli-Sforza, L. L. The Chinese human genome diversity project. *Proc. Natl Acad. Sci. USA* **95**, 11501–11503 (1998).
- Excoffier, L., Laval, G. & Schneider, S. Arlequin (version 3.0): an integrated software package for population genetics data analysis. *Evol. Bioinform. Online.* **1**, 47–50 (2005).
- Nei, M. *Molecular Evolutionary Genetics* (Columbia University Press, New York, NY, USA, 1987).
- Wright, S. *Evolution and the Genetics of Populations, Volume 4* (University of Chicago Press, Chicago, IL, USA, 1984).
- Peakall, R. O. D. & Smouse, P. E. Genalex 6: genetic analysis in Excel. Population genetic software for teaching and research. *Mol. Ecol. Notes* **6**, 288–295 (2006).
- Excoffier, L., Smouse, P. E. & Quattro, J. M. Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. *Genetics* **131**, 479–491 (1992).
- Smouse, P. E., Long, J. C. & Sokal, R. R. Multiple regression and correlation extensions of the Mantel test of matrix correspondence. *Syst. Biol.* **35**, 627 (1986).
- Smouse, P. E. & Long, J. C. Matrix correlation analysis in anthropology and genetics. *Am. J. Phys. Anthropol.* **35**, 187–213 (1992).
- Slatkin, M. A measure of population subdivision based on microsatellite allele frequencies. *Genetics* **139**, 457–462 (1995).
- Sinnott, R. W. Virtues of the Haversine. *Sky. Telescope* **68**, 158–159 (1982).
- Sagart, L., Blench, R. & Sanchez-Mazas, A. *The Peopling of East Asia: Putting Together Archaeology, Linguistics and Genetics* (RoutledgeCurzon, London, UK, 2005).
- Gordon, R. G. & Grimes, B. F. *Ethnologue: Languages of the World* (SIL International, Dallas, TX, USA, 2005).
- Pritchard, J. K., Stephens, M. & Donnelly, P. Inference of population structure using multilocus genotype data. *Genetics* **155**, 945–959 (2000).
- Falush, D., Stephens, M. & Pritchard, J. K. Inference of population structure using multilocus genotype data: dominant markers and null alleles. *Mol. Ecol. Notes* **7**, 574–578 (2007).
- Falush, D., Stephens, M. & Pritchard, J. K. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics* **164**, 1567–1587 (2003).
- Hubisz, M. J., Falush, D., Stephens, M. & Pritchard, J. K. Inferring weak population structure with the assistance of sample group information. *Mol. Ecol. Resour.* **9**, 1322–1332 (2009).
- Rosenberg, N. A. DISTRUCT: a program for the graphical display of population structure. *Mol. Ecol. Notes* **4**, 137–138 (2004).
- Karafet, T., Xu, L., Du, R., Wang, W., Feng, S., Wells, R. S. *et al*. Paternal population history of East Asia: sources, patterns, and microevolutionary processes. *Am. J. Hum. Genet.* **69**, 615–628 (2001).
- Hedrick, P. W. Perspective: highly variable loci and their interpretation in evolution and conservation. *Evolution* **53**, 313–318 (1999).
- Ding, Y. C., Wooding, S., Harpending, H. C., Chi, H. C., Li, H. P., Fu, Y. X. *et al*. Population structure and history in East Asia. *Proc. Natl Acad. Sci. USA* **97**, 14003–14006 (2000).
- Cavalli-Sforza, L. L. Genes, peoples, and languages. *Proc. Natl Acad. Sci. USA* **94**, 7719 (1997).
- Kutanun, W., Kampuansai, J., Colonna, V., Nakbunlung, S., Lertvicha, P., Seielstad, M. *et al*. Genetic affinity and admixture of northern Thai people along their migration route in northern Thailand: evidence from autosomal STR loci. *J. Hum. Genet.* **56**, 130–137 (2011).
- Kutanun, W., Kampuansai, J., Fuselli, S., Nakbunlung, S., Seielstad, M., Bertorelle, G. *et al*. Genetic structure of the Mon-Khmer speaking groups and their affinity to the neighbouring Tai populations in Northern Thailand. *BMC Genet.* **12**, 56 (2011).
- Su, B., Xiao, J., Underhill, P., Deka, R., Zhang, W., Akey, J. *et al*. 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 (1999).
- Wen, B., Li, H., Gao, S., Mao, X., Gao, Y., Li, F. *et al*. Genetic structure of Hmong-Mien speaking populations in East Asia as revealed by mtDNA lineages. *Mol. Biol. Evol.* **22**, 725–734 (2005).
- Chen, L. *Zhongguo Minzushi Gangyao (The Compendium of Chinese Nationality Histories)* (China Financial & Economic Press, Beijing, China, 1999).
- Ma, Y. *China's Minority Nationalities* (Foreign Languages Press, Beijing, China, 1989).

Supplementary Information accompanies the paper on Journal of Human Genetics website (<http://www.nature.com/jhg>)