European Journal of Human Genetics (2008) 16, 705–717

www.nature.com/ejhg

ARTICLE

A spatial analysis of genetic structure of human populations in China reveals distinct difference between maternal and paternal lineages

Fuzhong Xue[1,2], Yi Wang[1], Shuhua Xu[1,3], Feng Zhang[1], Bo Wen[1], Xuesen Wu[1], Ming Lu[1], Ranjan Deka[4], Ji Qian[1] and Li Jin*[,1,3]

[1]*MOE Key Laboratory of Contemporary Anthropology and Center for Evolutionary Biology, School of Life Sciences, Fudan University, Shanghai, China;* [2]*Department of Epidemiology and Biostatistics, School of Public Health, Shandong University, Jinan, China;* [3]*Department of Computational Genomics, CAS-MPG Partner Institute for Computational Biology, SIBS, CAS, Shanghai, China;* [4]*Department of Environmental Health, University of Cincinnati College of Medicine, Cincinnati, OH, USA*

Analyses of archeological, anatomical, linguistic, and genetic data suggested consistently the presence of a significant boundary between the populations of north and south in China. However, the exact location and the strength of this boundary have remained controversial. In this study, we systematically explored the spatial genetic structure and the boundary of north–south division of human populations using mtDNA data in 91 populations and Y-chromosome data in 143 populations. Our results highlight a distinct difference between spatial genetic structures of maternal and paternal lineages. A substantial genetic differentiation between northern and southern populations is the characteristic of maternal structure, with a significant uninterrupted genetic boundary extending approximately along the Huai River and Qin Mountains north to Yangtze River. On the paternal side, however, no obvious genetic differentiation between northern and southern populations is revealed.
European Journal of Human Genetics (2008) **16,** 705–717; doi:10.1038/sj.ejhg.5201998; published online 23 January 2008

Keywords: spatial genetic structure; maternal and paternal lineages; mitochondrial DNA; Y-chromosome; GIS; China

Introduction

Analyses of archeological, anatomical, linguistic, and genetic data suggested consistently the presence of a significant boundary between the populations of north and south in China.[1] Genetic differentiation between the southern and northern populations was observed at classic markers,[2–5] STR markers,[6] mtDNA,[7–9] and Y-chromosome SNP markers.[10,11] However, the exact location and the strength of this boundary have been remained controversial.[10,12] Using classic markers, Xiao *et al*[5] proposed a genetic boundary approximately located at Yangtze River. Wen *et al*[7] found that the mtDNA haplogroup distribution showed substantial differentiation between northern and southern Hans, whereas the differentiation of Y-chromosome haplogroups between the south and north is much more evasive. Furthermore, using three human genetic marker systems (mtDNA, Y-chromosome, autosomal STR) and one human virus, Ding *et al*[12] found that the north–south cline is virtually continuous and concluded that this can be better described by a model of simple isolation by distance. Therefore, the inconsistent and somewhat

***Correspondence: Professor L Jin, MOE Key Laboratory of Contemporary Anthropology and Center for Evolutionary Biology, School of Life Sciences, Fudan University, 220 Handan Road, Shanghai 200433, China. Tel: +021 5566 4382; Fax: +021 5566 4388;**
E-mail: ljin007@gmail.com
Received 30 January 2007; revised 4 December 2007; accepted 11 December 2007; published online 23 January 2008

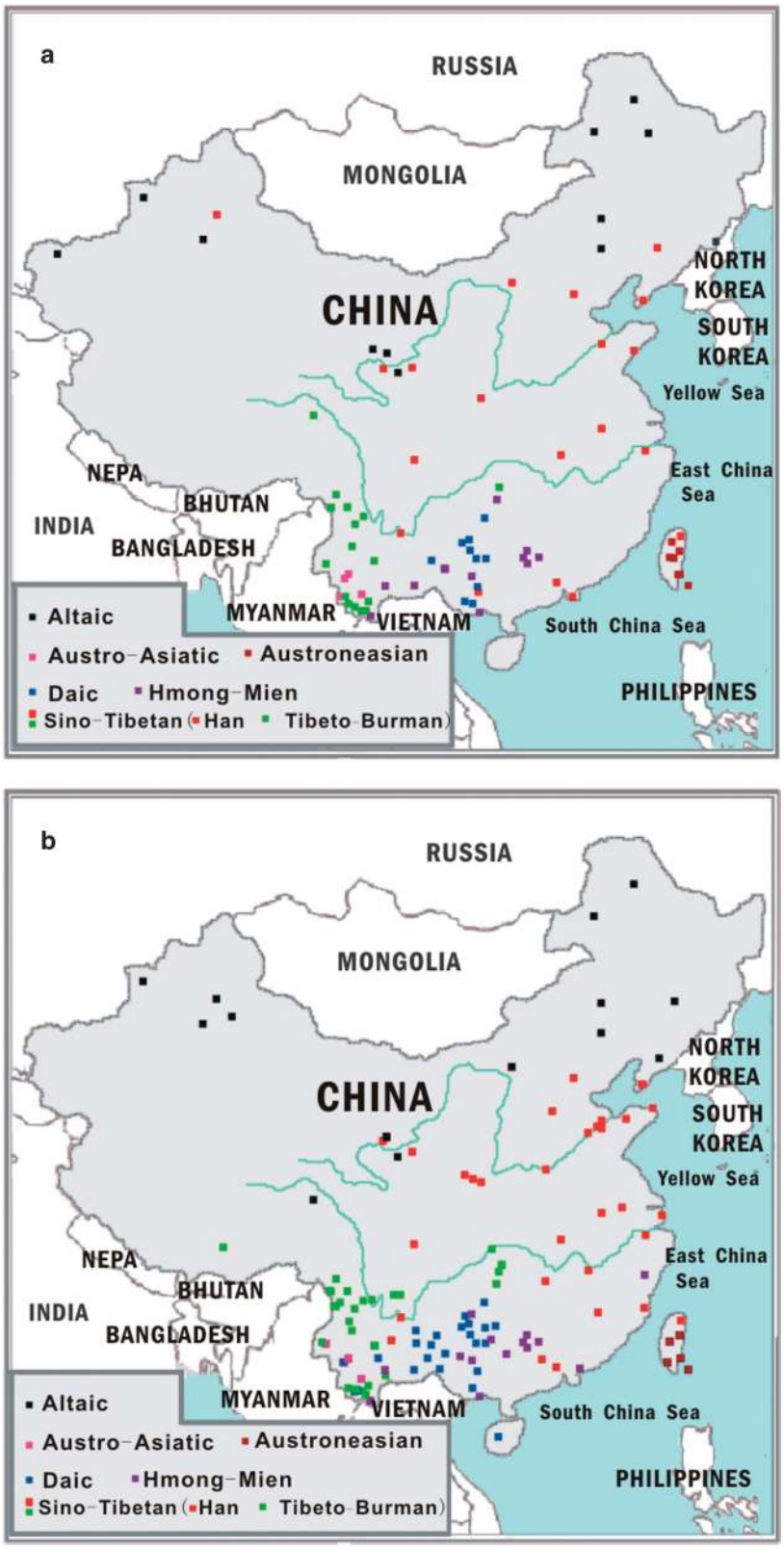

Figure 1 Locations of (**a**) 91 sampled populations for mtDNA and (**b**) 143 sampled populations for Y chromosome.

conflicting observations among different studies warrant a closer investigation of the spatial genetic structure of the populations in East Asia, especially considering the implication of such observations in understanding the origin and evolution of the populations and the application of such knowledge in designing molecular epidemiology studies. In this study, we systematically explored the spatial genetic structure and the boundary of north–south division of human populations in China.

To characterize the differentiation between northern and southern populations, especially the boundaries between them, the statistical technique should deal with both geographic locations of populations and their high-dimensional genetic data. The common statistic methods used in the aforementioned studies,[2,3,7–12] such as principle component analysis (PCA) and clustering analysis are less appropriate in reflecting spatial or geographic information.[13] In this paper, therefore, the PCA in combination with inverse distance-weighting (IDW) interpolation was used to visualize spatial genetic patterns and detect geographic genetic clines in mtDNA and Y-chromosome data.[14] In addition, the improved Monmomier's algorithm model[13,15] was used to identify the spatial genetic boundaries, and the genetic distograms were used to detect the statistical significance of spatial autocorrelation.[16] The geographic information system (GIS) is a powerful tool for management, analysis, and display of geographic information, it is used in this study to visualize spatial patterns on a map. It integrates common database operations and statistical analysis with unique visualization of geographic information offered by maps.

Mitochondrial DNA (mtDNA) and Y-chromosome polymorphisms have been studied extensively in the context of human population genetics,[1,17] which provide sufficient data for the analysis of spatial patterns of genetic structure. In this study, the spatial databases of 36 mtDNA haplogroups in 91 populations and 9 Y-chromosome haplogroups in 143 populations in China were developed, respectively. Such data were analyzed to characterize the spatial genetic structure and boundaries of genetic differentiation in human populations in China, with the emphasis on the comparison of such structure between the maternal and paternal lineages.

Material and methods

Samples and their spatial databases

Data on Y-chromosome and mtDNA of 3193 unrelated individuals from 80 Chinese populations speaking different languages across China were previously reported and included in this study.[7,18–20] Additional data were obtained from the literatures and added to this study, and the final sample sizes were expanded to 3435 individuals from 91 Chinese populations for mtDNA and 5790 individuals from 143 Chinese populations for Y-chromosome. These encompass the samples from all provinces in China. Figure 1 shows the locations of the samples, and a list describing the sources of the data is provided as supplements (see Supplementary Materials 1 and 2).

MtDNA and Y-chromosome polymorphisms and haplogroups

In the spatial databases, both the HVS motif and the coding region variations were used to define 36 haplogroups (*A*, *B**, *B*4, *B4a*, *B4b*1, *B5**, *B5a*, *B5b*, *C*, *D**, *D*5, *D5a*, *F**, *F1a*, *F1b*, *F1c*, *F2a*, *G*, *M**, *M7**, *M7a*, *M7a*1, *M7b**, *M7b*1, *M7b*2, *M7c*, *M8a*, *M9*, *N**, *N9a*, *R**, *R9a*, *R9b*, *R9c*, *Y*, and *Z*) following the phylogeny of East Asian mtDNA.[21] Thirteen bi-allelic Y-chromosome markers, *YAP*, *M*15, *M*130, *M*89, *M9*, *M*122, *M*134, *M*119, *M*110, *M*95, *M*88, *M*45, and *M*120 were used to define nine haplogroups (*C*, *D*, *F**, *K**, *O3**, *O3e*, *O*1, *O2a*, and *P**) following the Y-chromosome consortium nomenclature.[7]

Detection of geographic genetic clines

To quantify the spatial variance of mtDNA or Y-chromosome, the PCA in combination with IDW interpolation was used to visualize spatial genetic patterns and detecting geographic genetic clines.[14,22] PCA was first used to obtain principle component scores (PC1 and PC2) for each population. Then, the IDW algorithm was used to obtain synthetic maps of PC1 and PC2. IDW method has been used to create the contour maps of gene frequency distributions in human population genetics,[14] in which, for each point, its interpolated estimate was made based on values at nearby locations weighted by their distance from the point. In this paper, the Natural Breaks (Jenks) method was used to classify the geographic genetic clines.[23]

Identification of spatial genetic boundaries

Spatial boundaries indicate where abrupt changes were observed. In the present study, the 'improved Monmomier's algorithm' model (BARRIER version 2.2)[13,15,24] was used. The objective of Monmomier's algorithm is to visualize data contained in genetic distance matrix on a geographical map and to identify boundaries by finding the largest differences between pairs of neighboring samples (populations). Fst statistics was used as distance measure.

Detection of spatial genetic autocorrelation

To describe the spatial patterns for multiple haplogroups simultaneously, the genetic distogram analysis which was implemented in Spatial Genetic Software (SGS, version 1.0d) is used to detect the spatial genetic autocorrelation using Fst as genetic distance measure.[16] Genetic distograms represent graphs where mean genetic distances (Fst) between all pairs of population belonging to a spatial distance class were plotted against the spatial distance classes, the statistical significance of spatial genetic

autocorrelation are tested by a permutation procedure. To describe the spatial patterns for single haplogroup, the Moran's *I* statistic[25,26] together with the Moran Correlogram which were implemented in the software of CrimeStat III (version 3.0) are used to detect the spatial genetic autocorrelation for each haplogroup, the statistical significance of spatial genetic autocorrelation are tested by a Monte Carlo simulation procedure.[27]

All the maps in this study were created by arcGIS9.0 (Environmental Systems Research Institute Inc., USA).

Results

Maternal and paternal geographic genetic clines

Figure 2 shows the geographic genetic clines interpolated by PC1 and PC2 using mtDNA and Y-chromosome haplogroups. Figure 1a is the PC1 map (contributing proportion 19.83%), and Figure 1b is the PC2 map (14.84%) with 36 mtDNA haplogroups in 91 populations. The *PC*1 map reveals an obvious north–south geographic genetic cline, whereas the PC2 map reveals a west–east cline. The Figure 1c is the PC1 map (33.07%), and Figure 1d is the PC2 map (17.07%), for 9 Y-chromosome haplogroups in 143 populations. The north–south cline for Y-chromosome is much less pronounced. Therefore, there are different geographic genetic cline patterns between maternal and paternal lineages.

Maternal and paternal spatial genetic boundaries

When all Han and non-Han populations are included, for both mtDNA and Y-chromosome data, genetic boundaries are mainly located in the peripheral mountainous regions (Figure 3a and c). We failed to observe statistically significant genetic boundaries between north and south.

However, when only Han populations are included, genetic boundaries between the northern and southern populations start to emerge (Figure 3b and d). Such division is statistically significant with the maternal lineages, but much weaker with the paternal lineages. On the maternal side, there are significant uninterrupted

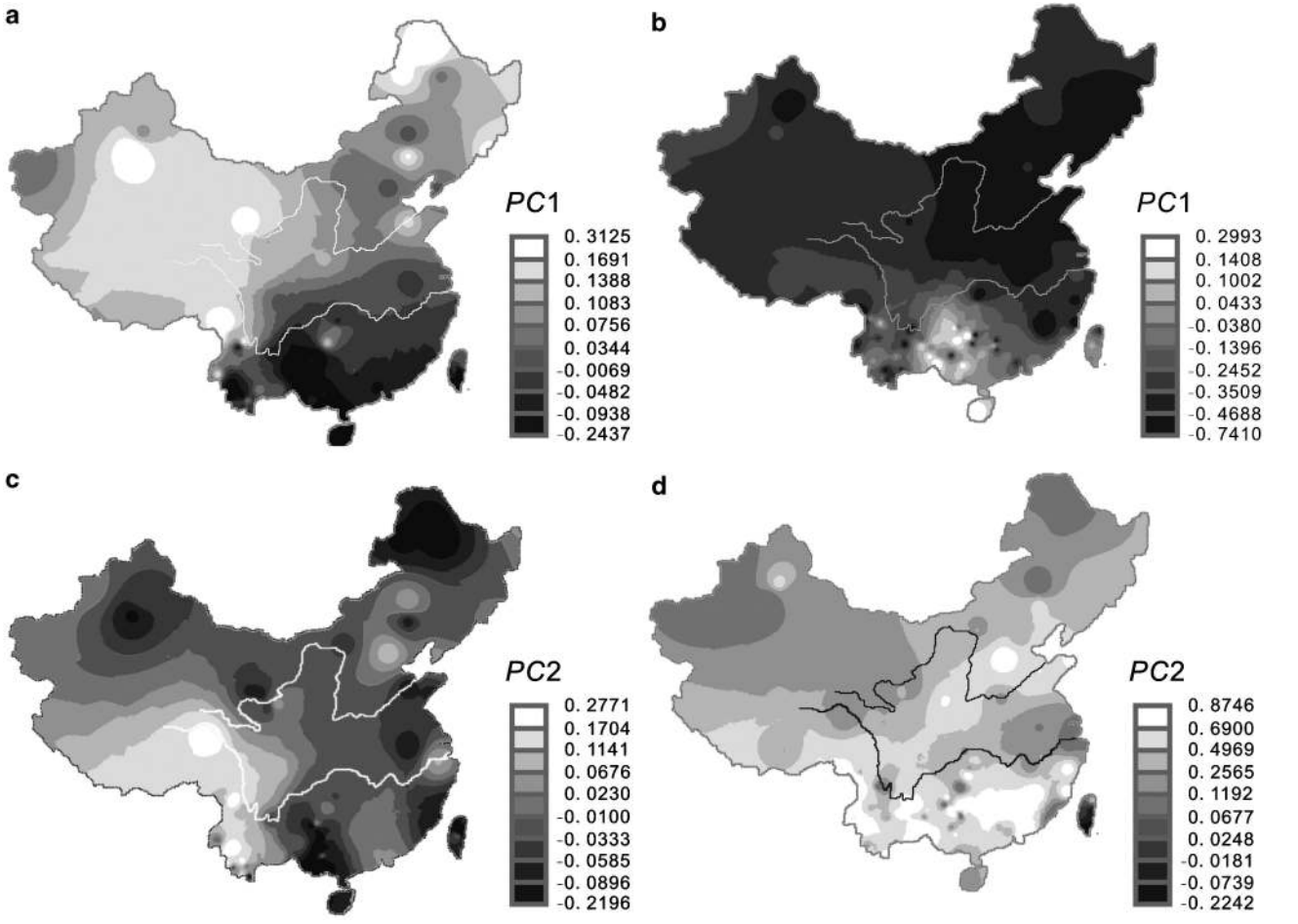

Figure 2 The geographic genetic clines interpolated using (**a**) *PC*1 (19.83%) and (**b**) *PC*2 (14.84%) with 36 mtDNA haplogroups in 91 populations respectively. (**c**) and (**d**) The maps of *PC*1 (33.07%) and *PC*2 (17.07%) with 9 haplogroups in 143 populations, respectively, are shown.

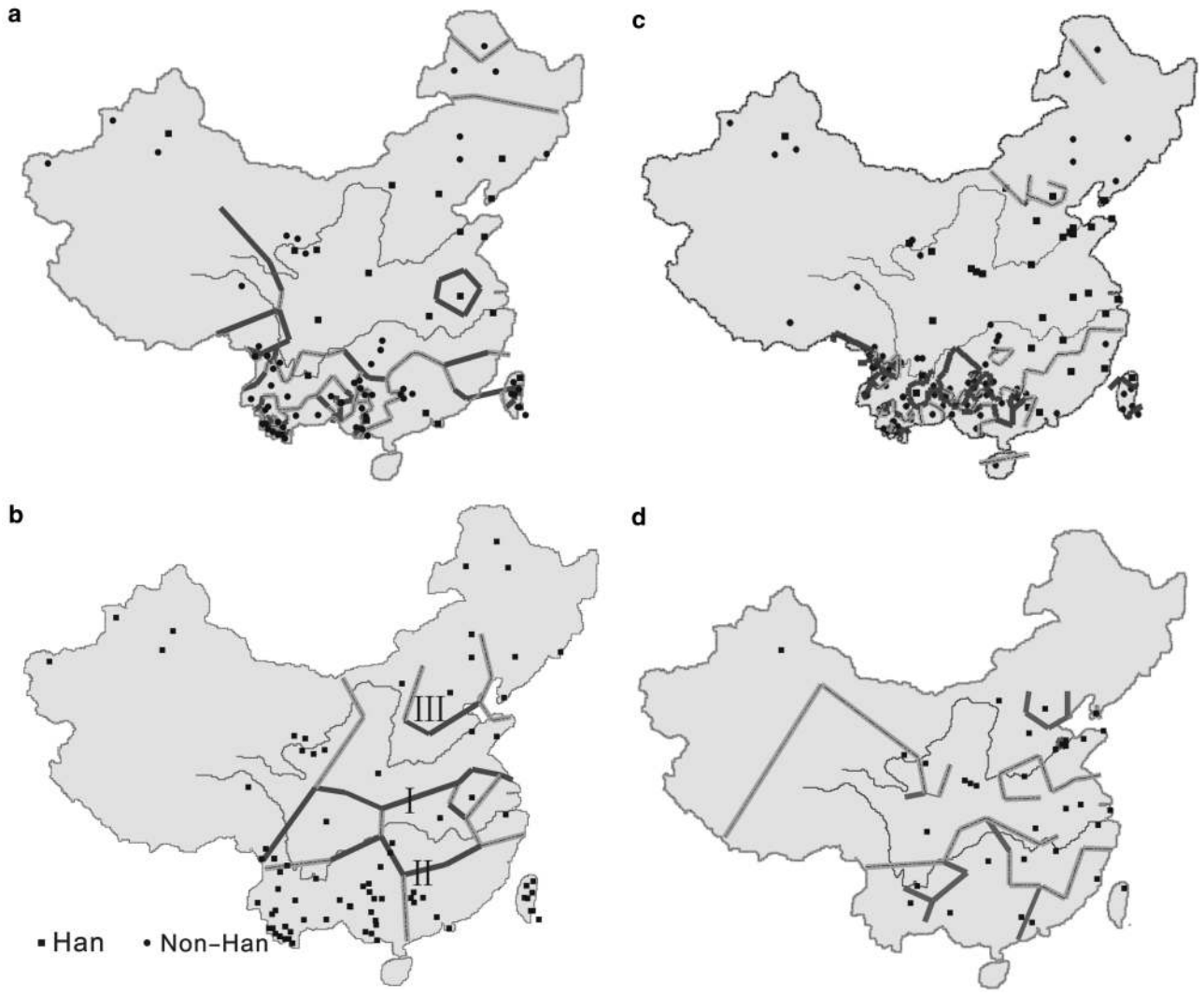

Figure 3 Spatial genetic boundaries of mtDNA and Y-chromosome in Chinese populations. Boundaries in (**a**) were calculated using 36 mtDNA haplogroups using 91 Han and no-Han populations, whereas boundaries in (**b**) were calculated using 36 mtDNA haplogroups in 19 Han populations. Boundaries in (**c**) were calculated by 9 Y-chromosome haplogroups using 143 Han and no-Han populations, whereas boundaries in (**d**) are calculated by 9 Y-chromosome haplogroups only using 35 Han populations. ▬ ▬ are spatial genetic boundaries with thickness of each edge proportional to its bootstrap score. ▬ is the score *that is* greater than 80%, whereas ▬ denotes the score that is less than 80%.

genetic boundaries between the populations from the north and the south, with the most prominent division extending approximately along the Huai River and Qin Mountains that are north to Yangtse River (Figure 3b). Two other boundaries are also observed with the one south to Yangtse River and the other north to Yellow River, although their statistical importance is much less significant than the one mentioned earlier based on their respective bootstrap values (Figure 3b). On the paternal side, the presence of genetic boundaries reveals a completely different pattern from their maternal counterparts. Boundaries are observed but in much *more fragmented way*. There are uninterrupted genetic boundaries between north and south, but they are not statistically significant (see Figure 3d), indicating the northern and southern populations are less differentiated at paternal lineages than they are at maternal lineages.

Spatial autocorrelation of maternal and paternal genetic structure

Genetic distogram analysis is used to examine the statistical significance of spatial autocorrelation for multiple haplogroups simultaneously. Figure 4 shows the distograms of average Fst in 14 spatial distance classes calculated based on the frequencies of mtDNA and Y-chromosome haplogroups, respectively. In these graphs, mean genetic distances are plotted against geographic distances. The lines for 95% confidence interval (CI) are

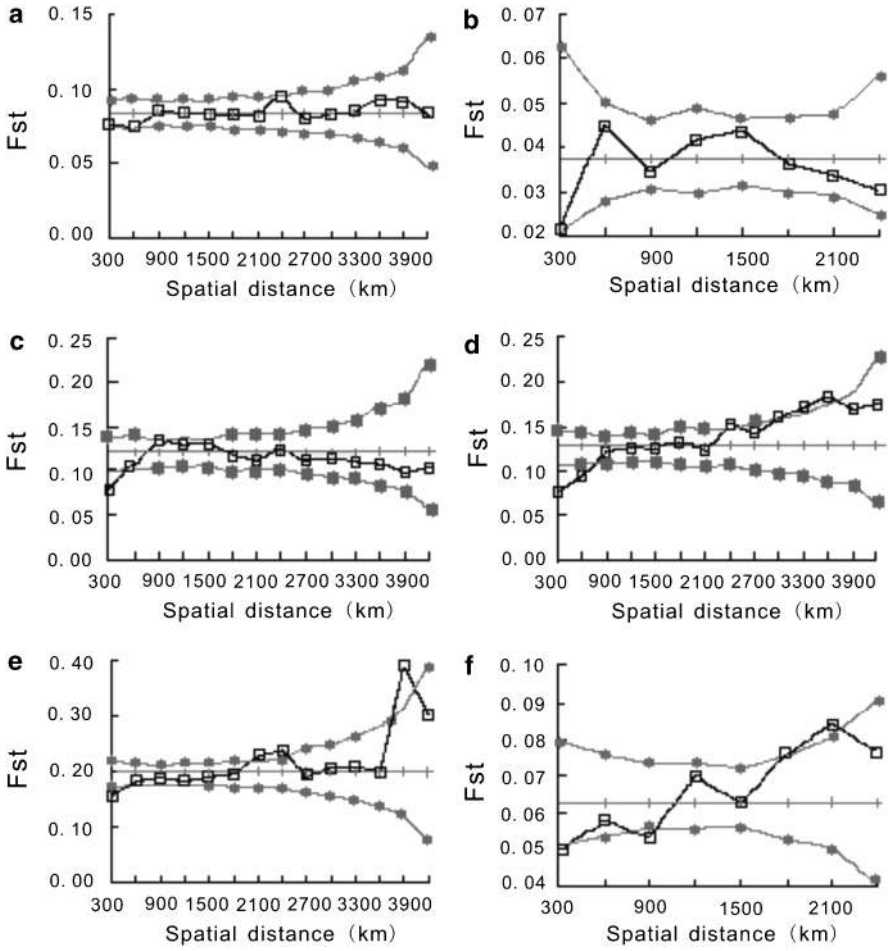

Figure 4 The distograms of average *Fst* in 14 spatial distance classes calculated by the frequencies of mtDNA and Y-chromosome haplogroups. The lines include the 95% CI of 1000 permutations (□: observed, +: reference/mean, ■: lower and upper 95% CI). (**a**) The distogram of mtDNA calculated by 36 haplogroups in 91 Han and non-Han populations is shown; (**b**) the distogram of mtDNA calculated by 36 haplogroups only in 19 Han populations is shown; (**c**) and (**d**) the distograms of mtDNA calculated by 10 NDHs and by 23 SDHs in 91 Han and non-Han populations, respectively, are shown. (**e**) The distogram of Y-chromosome calculated by 9 haplogroups in 143 Han and non-Han populations is shown; (**f**) the distogram of Y-chromosome calculated by 9 haplogroups only in 35 Han populations is shown.

obtained based on 1000 permutations. Again, genetic distograms reveal different spatial structure between maternal and paternal lineages.

On the maternal side, spatial autocorrelation is found neither in 91 Han and non-Han populations (Figure 4a); nor in 19 Han populations (Figure 4b). This indicates that there are some maternal sub-structures with genetic differentiation distributing stochastically in China. On the paternal side, genetic distance increases with geographic distance (Figure 4e and f). Genetic distances are significantly higher in the classes ranging from 1800 to 2100 km, indicating that there is a substantial paternal spatial autocorrelation across landscape from north to south in China.

To further test the spatial autocorrelation for each haplogroup of mtDNA and Y-chromosome, the Moran's *I* statistic together with the Moran correlogram are used to detect the spatial genetic autocorrelation for each mtDNA or Y-chromosome haplogroup. Supplementary Tables 1 and 2 show the Moran's *I* for 91 Han and non-Han populations and for 19 Han populations, respectively, in 11 spatial distance classes calculated based on the frequency of each mtDNA haplogroup. Supplementary Tables 3 and 4 show the Moran's *I* for 143 Han and non-Han populations and for 35 Han populations, respectively, in 11 spatial distance classes calculated based on the frequency of each Y-chromosome haplogroup. The statistical significance of spatial genetic autocorrelation for each haplogroup are tested by 1000 Monte Carlo simulations. Similarly, Moran's *I*'s in Moran correlogram reveal different spatial structure between maternal and paternal lineages.

On the maternal side, when all Han and non-Han populations are included, 21 haplogroups (A, D*, D5a, G, M9, Z, M*, B4a, B5a, F*, F1a, M7*, M7b*, M7b1, B*, B5*,

F1b, F2a, N9a, R9c, and Y) present spatial autocorrelation, other 15 haplogroups (C, D5, M7c, M8a, N*, B4b, B4b1, R9a, R9b, B4, B5b, F1c, M7a, M7b2, and R*) do not present spatial autocorrelation (Supplementary Table 1). This indicates that some mtDNA haplogroups present their substantial spatial autocorrelation in the local geographic regions, whereas others are distributing stochastically across landscape of China. However, the synthetical maternal spatial pattern with multiple mtDNA haplogroups simultaneously present no spatial autocorrelation across landscape of China (Figure 4a), indicating that there are some maternal sub-structures with genetic differentiation distributing stochastically in global geographic regions. When only Han populations are included, most mtDNA haplogroups present no spatial autocorrelation except D*, N*, F1a, and M7b* (Supplementary Table 2), indicating that most mtDNA haplogroups are distributing stochastically in Han populations across landscape from north to south in China.

On the paternal side, when all Han and non-Han populations are included, most Y-chromosome haplogroups (C, D, F*, K*,O3e, O1, O2a, and P*) present spatial autocorrelation except O3* (Supplementary Table 3), indicating that most Y-chromosome haplogroups are not distributing stochastically in populations across landscape from north to south in China. Again, when only Han populations are included, most Y-chromosome haplogroups also present spatial autocorrelation except D and P* (Supplementary Table 4), indicating that there is a substantial paternal spatial autocorrelation in Han populations across landscape from north to south in China.

Spatial genetic distribution of maternal lineages

Kivissild *et al*[21] determined the phylogenetic backbone of the East Asian mtDNA tree. Their results confirm that the East Asian mtDNA lineages are region-specific and completely covered by the two superhaplogroups *M* and *N*. The phylogenetic partitioning based on complete mtDNA sequences corroborates existing RFLP-based classification of Asian mtDNA types and supports the distinction between northern and southern populations.[21]

Figure 5 shows the frequency maps of haplogroups *M* (including its 16 sub-haplogroups) and haplogroups *N* (including its 20 sub-haplogroups), respectively. Each map of haplogroup is created based on its frequencies in 91 populations (see Figure 5a and b). The tree is rooted using haplogroup L3 as an outgroup, and it has two major branches (*M* and *N*). The maps show that distribution of mtDNA haplogroups presents a distinct north–south differentiation in China. The frequency of haplogroups *M* is much higher in the north encompassing Northern Han, Altaic, and northern Tibetan-Burman populations (Figures 1a and 5a), whereas the frequency haplogroups *N* is much higher in the south encompassing Southern Han, Daic, Hmong-Mien, Austro-Asiatic, Austroneasian, and southern Tibetan-Burman populations (see Figures 1a and 5a). Using boundary I in Figure 3b, most haplogroups can be classified to either southern dominating haplogroup (SDH including R, B, R9, F, and M7) or northern dominating haplogroup (NDH including A, N, M9, D, G, M8, and CZ) based on its frequency distribution, although several haplogroups cannot be classified to SDH or NDH.

Table 1 shows the distribution of northern and southern dominating haplogroups of mtDNA. Haplogroups A, *C*, *D**, *D5*, *D5a*, *G*, *M7c*, *M8a*, *M9*, *N**, and *Z* are identified as NDH, with much higher frequencies in north than in south significantly. Haplogroups, *M**, *B**, *B4*, *B4a*, *B4b1*, *B5**, *B5a*, *F**, *F1a*, *F1b*, *F1c*, *F2a*, *M7**, *M7a*, *M7b**, *M7b1*, *M7b2*, *R**, *R9a*, *R9b*, and *R9c* are identified as SDH, with much higher frequencies in south than in north. Most of the major haplogroups derived from *M* lineage are NDH except for M7, whereas most of the major haplogroups derived from *N* lineage are SDH except for N9 and A.

To investigate further the spatial genetic structures of NDHs and SDHs, their synthetic maps (Figure 5c and d) and their distograms (Figure 4c and d) are created, respectively, for NDHs and SDHs. Figure 5c and d are the synthetic maps of mtDNA by PC1 (30.33%) with 10 NDHs and by PC1 (22.50%) with 23 SDHs in 91 populations, respectively. Figure 4c and d are the distograms of mtDNA calculated by 10 NDHs and by 23 SDHs in 91 populations, respectively. In the maps for NDHs and SDHs, the northern genetic structure and the northern genetic structure become quite distinct, and the dissimilarity between northern and southern populations becomes more pronounced, especially between Northern Hans and Southern Hans (Figure 5c and d). When the distograms were calculated by 10 NDHs (Figure 4c) or by 23 SDHs, (Figure 4d) respectively, significant spatial autocorrelation were detected. For NDHs, genetic distances are significantly lower than that expected by chance in the first distance classes (up to 300 km) and significantly higher in the classes of 900 km. For SDHs, genetic distance increases with geographic distance. Genetic distances are significantly lower than that expected by chance in the first two distance classes (up to 600 km) and significantly higher in the classes ranging from 2400 to 3600 km. This indicates that there are substantial maternal spatial autocorrelation in north and south.

Spatial genetic distribution of paternal lineages

The phylogeny of Y-chromosome haplogroups in East Asians was obtained following Jin and Su.[1] It was rooted using haplogroup *M*168, and was divided into three major branches (*M*89, *M*1, and *M*130). Figure 6 shows the frequency maps representing the major branches M89 and the haplogroup K* of the phylogeny of Y-chromosome, and each haplogroup map is created based on its frequency in 143 populations. As expected, the spatial pattern of

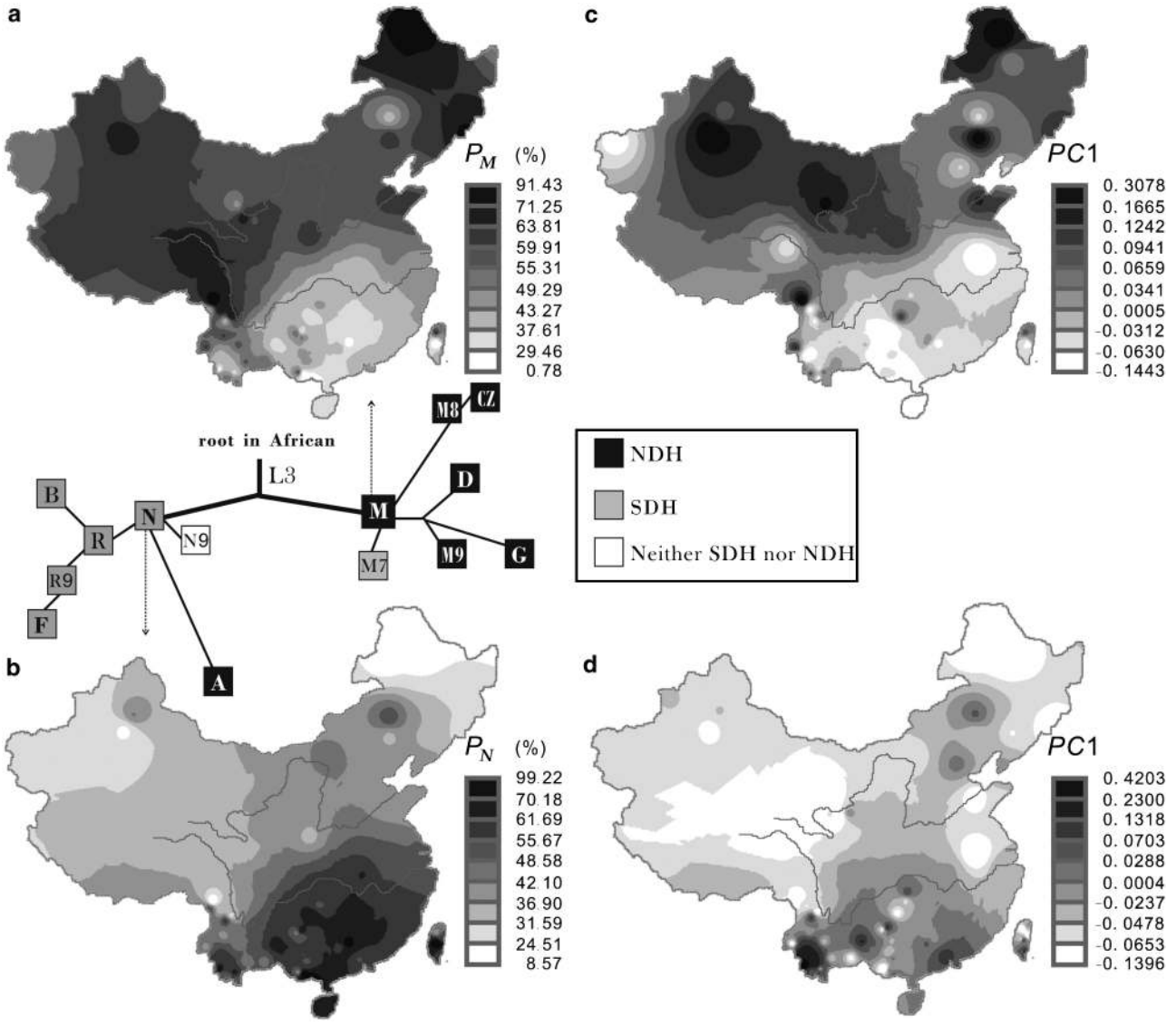

Figure 5 The frequency maps of dominating mtDNA haplogroups. (**a**) The frequency map of haplogroup *M* (including 16 sub-haplogroups) is shown, and (**b**) the frequency map of haplogroup *N* (including 20 sub-haplogroups) is shown. Each map of haplogroup was created based on its frequency in 91 populations. The tree was rooted using haplogroup *L*3 as an outgroup, and it has two major branches (*M* and *N*). (**c**) and (**d**) The synthetic maps of mtDNA by *PC*1 (30.33%) with 10 NDHs and by *PC*1 (22.50%) with 23 SDHs in 91 populations, respectively, are shown.

Y-chromosome haplogroups is quite different from mtDNA haplogroups.

The major branch M89 is prevalent in all populations sampled without significant differentiation between north and south (Figure 6a). Furthermore, the sub-branch M9 along with its descendent sub-branch groups (M95, M119, and M122) are also prevalent in all populations with very high frequencies in most populations except for Altaic populations and Tibetan. The distribution of haplogroup K* demarcates the outline of Tibetan-Burman corridor extending into Yunnan (Figure 6b). Therefore, most of the haplogroups cannot be classified into NDH or SDH. However, the frequency differences were indeed observed among linguistic groups, suggesting a correlation between Y-chromosome haplogroups with linguistic classification. For examples, haplogroups O3*, O3e, and K* have much higher frequencies in most populations except in Austroneasian population, while haplogroups C, P*, and F* in Altaic, haplogroup O1 in Austroneasian, haplogroup O2a in Daic, and haplogroup D in Tibeto-Burman are much dominating, respectively (Table 2).

To identify the genetic homogeneity between north and south population in paternal lineage, the Han populations for Y-chromosome were divided into north Han population and south Han population using boundary *I* of mtDNA in Han population (Figure 3b), and then we tested the

Table 1 Distribution of northern and southern dominating haplogroups of mtDNA

	Frequencies in south		*Frequencies in north*			
Haplogroup	n	*Frequency (95% CI) (%)*	n	*Frequency (95% CI) (%)*	*Fisher's exact text (P-value)*	*Dominating types*
A	128	5.48 (4.56, 6.41)	90	8.17 (6.56, 9.79)	**0.0034**	NDH
C	90	3.86 (3.07, 4.64)	72	6.54 (5.08, 8.00)	**0.0007**	NDH
*D**	284	12.17 (10.84, 13.49)	225	20.44 (18.05, 22.82)	**0.0000**	NDH
D5	59	2.53 (1.89, 3.16)	47	4.27 (3.07, 5.46)	**0.0079**	NDH
D5a	25	1.07 (0.65, 1.49)	25	2.27 (1.39, 3.15)	**0.0088**	NDH
G	58	2.49 (1.85, 3.12)	65	5.90 (4.51, 7.30)	**0.0000**	NDH
M7c	25	1.07 (0.65, 1.49)	23	2.09 (1.33, 3.12)	**0.0279**	NDH
M8a	153	6.56 (5.55, 7.56)	137	12.44 (10.49, 14.39)	**0.0000**	NDH
M9	21	0.90 (0.52, 1.28)	25	2.27 (1.39, 3.15)	**0.0021**	NDH
*N**	28	1.20 (0.76, 1.64)	39	3.54 (2.45, 4.63)	**0.0000**	NDH
Z	26	1.11 (0.69, 1.54)	39	3.54 (2.45, 4.63)	**0.0000**	NDH
*M**	240	10.28 (9.05, 11.51)	84	7.63 (6.06, 9.19)	**0.0124**	SDH
B4a	153	6.56 (5.55, 7.56)	23	2.09 (1.33, 3.12)	**0.0000**	SDH
B4b1	62	2.66 (2.00, 3.31)	11	1.00 (0.50, 1.78)	**0.0014**	SDH
B5a	142	6.08 (5.11, 7.05)	9	0.82 (0.37, 1.55)	**0.0000**	SDH
*F**	80	3.43 (2.69, 4.17)	19	1.73 (1.04, 2.68)	**0.0044**	SDH
F1a	250	10.71 (9.46, 11.97)	40	3.63 (2.61, 4.91)	**0.0000**	SDH
*M7**	24	1.03 (0.62, 1.44)	3	0.27 (0.06, 0.79)	**0.0210**	SDH
*M7b**	85	3.64 (2.88, 4.40)	12	1.09 (0.56, 1.90)	**0.0000**	SDH
M7b1	101	4.33 (3.50, 5.15)	25	2.27 (1.47, 3.33)	**0.0025**	SDH
R9a	43	1.84 (1.30, 2.39)	4	0.36 (0.01, 0.72)	**0.0002**	SDH
R9b	34	1.46 (0.97, 1.94)	3	0.27 (0.06, 0.79)	**0.0011**	SDH
*B**	16	0.69 (0.35, 1.02)	2	0.18 (0.02, 0.65)	0.0740	SDH
B4	94	4.03 (3.23, 4.83)	31	2.82 (1.92, 3.97)	0.0794	SDH
*B5**	10	0.43 (0.16, 0.69)	3	0.27 (0.06, 0.79)	0.5684	SDH
B5b	15	0.64 (0.32, 0.97)	16	1.45 (0.83, 2.35)	0.0513	
F1b	54	2.31 (1.70, 2.93)	25	2.27 (1.47, 3.33)	1.0000	SDH
F1c	18	0.77 (0.42, 1.13)	11	1.00 (0.50, 1.78)	0.5496	SDH
F2a	30	1.29 (0.83, 1.74)	11	1.00 (0.50, 1.78)	0.6137	SDH
M7a	3	0.13 (0.03, 0.38)	1	0.09 (00.00, 0.51)	1.0000	SDH
M7b2	9	0.39 (0.13, 0.64)	9	0.82 (0.37, 1.55)	0.1274	SDH
N9a	50	2.14 (1.55, 2.73)	29	2.63 (1.77, 3.76)	0.3936	
*R**	20	0.86 (0.48, 1.23)	5	0.45 (0.15, 1.06)	0.2814	SDH
R9c	9	0.39 (0.13, 0.64)	0	0.00 (0.00, 0.27)	0.0656	SDH
Y	9	0.39 (0.13, 0.64)	8	0.73 (0.31, 1.43)	0.1978	

Note: The significant level is defined as $P<0.05$. Although frequency of haplogroups *B**, *B4*, *B5**, *F1b*, *F1c*, *F2a*, *M7a*, *M7b2*, *R**, and *R9c* are not statistical significant between south and north, they are in fact dominating in the language groups in south. Thus, they were identified as SDH.

difference for each haplogroup of Y-chromosome between north Han population and south Han population. Table 3 shows the distribution of Y-chromosome haplogroups in northern Han and southern Han populations. It indicates that southern Hans and northern Hans share similar frequencies of Y-chromosome haplogroups (Table 3), which are characterized by carrying the *M*89, O3*, O3e, and K* mutations that are prevalent in almost all Han populations studied ($P>0.05$). Haplogroups C and D, whose frequencies are not prevalent in most Han populations, are also *not a significant* difference between southern Hans and northern Hans populations respectively ($P>0.05$). Although the difference of haplogroups F*, O1, O2a, and P* between southern Hans and northern Hans are significant, respectively ($P<0.05$), they are infrequent in most Han populations except that the haplogroup O1 presents a higher frequency in southern Hans (14.09%, 11.40–17.15%). Therefore, the paternal lineage is different from the maternal lineage, most haplogroups of Y-chromosome can be classified to neither SDH nor NDH.

Discussion

Spatial genetic structure: maternal *versus* paternal

For maternal lineages, we show that (1) there is a distinct north–south geographic genetic cline (Figure 2a), (2) there is a substantial genetic differentiation between northern and southern populations (Figure 2a), and (3) there is an identifiable boundary dividing the northern and southern populations (Figure 3b). It should be noted that the boundary dividing the south and north emerges only when non-Han populations are excluded (Figure 3b). When all populations are analyzed, the boundaries are mainly located in the peripheral regions of China where minority nationalities reside, although largely not significant, (Figure 3a). The most prominent division extends

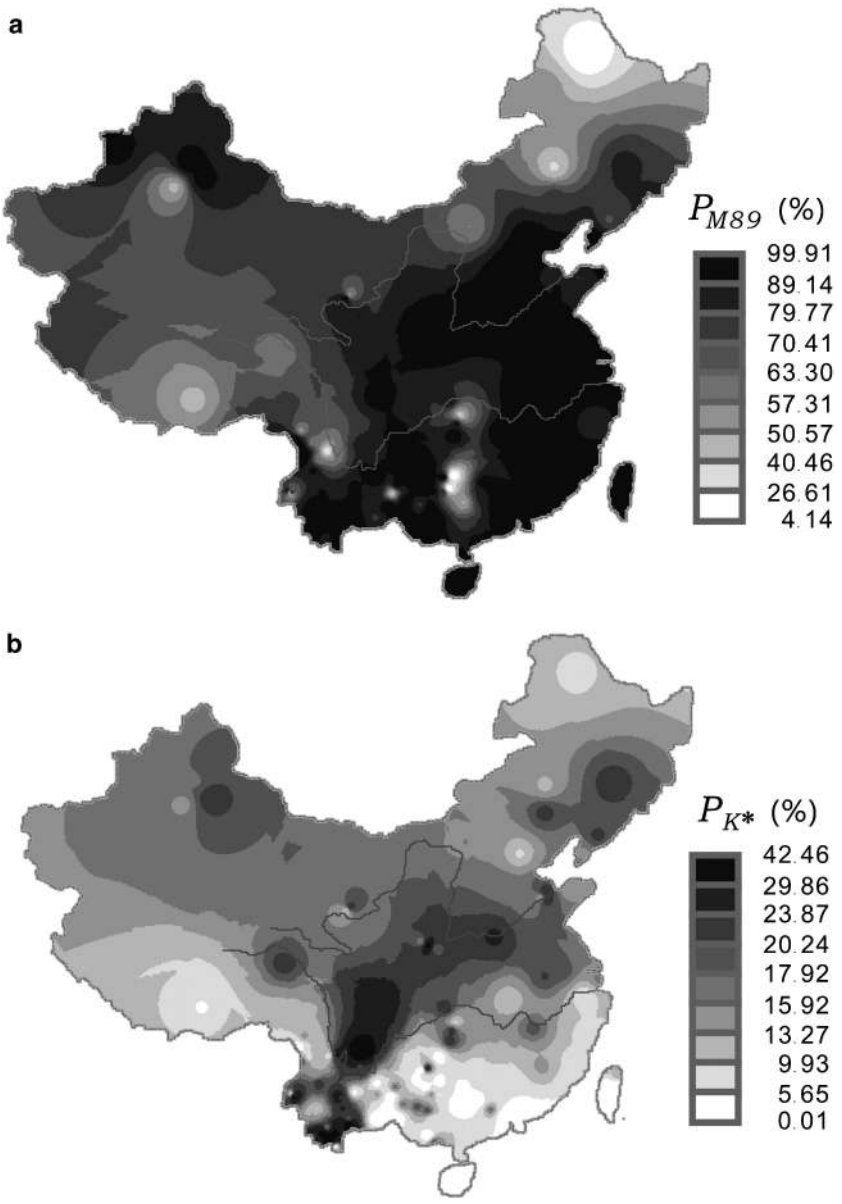

Figure 6 Two frequency maps of Y-chromosome haplogroups in 143 Han and non-Han populations. (**a**) The map of the major branch M89 with 7 haplogroups is shown, and (**b**) the map of haplogroup K* is shown.

approximately along the Huai river and Qin mountains that are north to Yangtse river (Figure 3b), inconstant with what was proposed by Xiao *et al*[5] using classic markers.

To delineate the geographic distribution of mtDNA haplogroups, their frequency maps are created using spatial data of 36 mtDNA haplogroups in 91 populations following the backbone of the mtDNA phylogeny.[21] The branch *M* is primarily distributed in the north whereas branch *N* in the south with a few important exceptions (Figure 5a and b). In five major lineages derived from *M*, four of them (*M*8, *M*9, *G*, and *D*) are primarily distributed in the north, but the *M*7 including its sub-branches are primarily distributed in Daic populations (Figure 5), a group of southern natives in Southeast Asia where was the entry point of modern humans in East Asia.[1,6,9,10,28] In three major lineages derived from *N*, the branch *R* including their sub-branches is primarily distributed in north, but the A is primarily distributed in north Tibeto-Burman, the branch *N9* is distributed in all over East Asia (Figure 5; Table 1).

Each mtDNA haplogroup is classified to either SDH or NDH based on its frequency distribution (Table 1). In the spatial genetic structures using either NDHs or SDHs, the northern genetic structure and the northern genetic structure become quite distinct, and the distinction

Table 2 Y-chromosome haplogroup frequencies in different population

						Sino-Tibetan	
Haplogroup	*ALT*	*A-A*	*AUS*	*DAC*	*H-M*	*HAN*	*TB*
C	**25.36**	9.33	0.00	5.12	5.21	7.30	6.60
D	6.30	0.00	0.00	4.12	3.87	1.81	**12.69**
*F**	**11.75**	3.30	0.00	3.50	0.89	5.98	5.09
*K**	17.19	15.94	0.95	8.24	10.52	16.24	19.25
*O3**	10.89	20.33	5.21	10.49	35.86	30.66	21.74
O3e	12.18	23.63	1.42	12.36	17.56	23.81	22.16
O1	0.21	0.00	**82.94**	9.12	5.70	6.91	3.53
O2a	0.29	27.84	9.48	**46.82**	20.39	3.51	7.65
*P**	**14.04**	0.00	0.00	0.25	0.00	3.79	1.28
Sample size	698	182	211	801	672	1823	1403

Note: ALT, Altaic; A-A, Austro-Asiatic; AUS, Austroneasian; DAC, Daic; H-M, Hmong-Mien; TB, Tibeto-Burman.

Table 3 Distribution of Y-chromosome haplogroups in northern Han and southern Han populations

	Frequencies in south		*Frequencies in north*		
Haplogroup	n	*Frequency (95% CI) (%)*	n	*Frequency (95% CI) (%)*	*Fisher's exact text (P-value)*
M89	548	91.95 (89.46, 94.00)	879	89.24 (87.13, 91.11)	0.0807
C (*M130*)	40	6.71 (4.84, 9.02)	84	8.53 (6.86, 10.45)	0.2102
D (*M1*)	8	1.34 (0.58, 2.63)	22	2.23 (1.40, 3.36)	0.2554
*F**	8	1.34 (0.58, 2.63)	75	7.61 (6.04, 9.45)	**0.0000**
*K**	93	15.60 (12.78, 18.77)	174	17.67 (15.33, 20.19)	0.2996
*O3**	166	27.58 (24.28, 31.64)	291	29.54 (26.71, 32.50)	0.4924
O3e	144	24.16 (20.77, 27.80)	252	25.58 (22.88, 28.43)	0.5495
O1 (*M119*)	84	14.09 (11.40, 17.15)	31	3.15 (2.15, 4.44)	**0.0000**
O2a (*M95*)	45	7.55 (9.56, 9.97)	14	1.42 (0.77, 2.37)	**0.0000**
*P** (*M45*)	8	1.34 (0.58, 2.63)	42	4.26 (3.09, 5.72)	**0.0010**
Total	596		985		

The significant level is defined as $P<0.05$.

between northern and southern populations becomes more prominent, especially between Northern Hans and Southern Hans (Figures 4c,d and 5c,d). Most mtDNA haplogroups are distributing stochastically in Han populations (Supplementary Table 2), and there are some maternal sub-structures with genetic differentiation distributing stochastically in Han populations (Figure 4b).

The paternal spatial genetic structure reveals a completely different pattern from what are observed in the maternal lineages. Unlike mtDNA, no obvious genetic differentiation between northern and southern populations is observed on the paternal side, even when only Han populations are included (Table 3). When all populations are included in the analysis, significant uninterrupted boundaries are observed in the peripheral regions separating Han populations and their nearby minority nationalities. When only Han populations are included, there is an absence of significant uninterrupted paternal genetic boundaries between Northern Hans and Southern Hans (Figure 3c and d); most Y-chromosome haplogroups present their substantial spatial autocorrelation between Han populations (Supplementary Table 4); and there is a substantial paternal spatial autocorrelation across landscape from north to south in China (Figure 4f).

In the past two millennia, there have been major population movements toward the south in China.[8,29–32] In particular, Wen *et al*[7] showed that such movements were sex-biased and mostly involving much more males than the females. These sex-biased gene flows, therefore, constituted a great deal of impact on the genetic structures of the extant populations and led to the differential structures of the populations between the maternal and paternal lineages as seen in this study.

Spatial pattern of genetic boundaries: Han *versus* Han and non-Hans

For both maternal and paternal lineages, genetic boundaries between the northern and southern populations start to emerge when only Han populations are included in the analysis (Figure 3b and d). This indicates that the patterns of spatial genetic boundaries are scale-dependent. The Fst values between Han and the populations in the Southwest China are much higher than those between Hans. When all non-Han populations are removed from the analysis, the Fst values between south Hans and north Hans become pronounced, and genetic boundaries between the northern and southern populations emerged (Figure 3b and d). Such scale-dependent effect was also observed in a study of

phylogenetic relationship of the populations within and around Japan using 105 short tandem repeat polymorphic loci.[33]

Spatial database and statistical methods

Although our spatial database encompasses 3435 individuals (91 Chinese populations) for mtDNA and 5790 individuals (143 Chinese populations) for Y-chromosome (Figure 1), the distribution and density of the sample points or populations are far from satisfactory given the complexity of the genetic structure in East Asia. On the other hand, as the level of resolution for the mtDNA is higher than the one presented by the Y-chromosome (only nine wide haplogroups are analyzed for the Y-chromosome whereas 36 haplogroups are analyzed for the mtDNA). There could be bias in the results of differentiation between maternal and paternal lineages. Another drawback of this study, which may have compromised the accuracy of the results, is the exclusion of the data from other important areas in East Asia and Southeast Asia, largely due to the lack of research effort on these populations.

Many approaches can be used for creating interpolated contour maps of genetic variables: Cavalli-Sforza method in Genography,[4] IDW method,[14] and the Kriging technique.[34] In the present study, we chose IDW algorithms for displaying spatial genetic patterns and for detecting geographic genetic clines, since it basically generates similar to or slightly better results than those who use other methods by comparing their results with each other using the data in this study (data not shown).

Several approaches can be used for detecting spatial genetic boundaries, such as *Wombling*, spatial analysis of molecular variance (SAMOVA), and the improved Monmomier's algorithm method.[13,15] We chose the improved Monmomier's algorithm (BARRIER version 2.2),[13,15] for identifying the spatial genetic boundaries, since it avoids potential artificial continuities or discontinuities in interpolation of the landscape in *Wombling*, and works slightly better than SAMOVA in finding spatial genetic boundaries.[13,15,35]

Different statistics can be used to detect the spatial genetic autocorrelation. Moran's index and Geary's index are among the most frequently used measures.[25,26] More recently multi-locus measures of spatial autocorrelation based on genetic distances were introduced, and a new statistics, called genetic distograms, has been created to detect spatial genetic autocorrelation, and to test the statistical significance of spatial autocorrelation.[36] In the present study, we choose genetic distograms that was implemented in Spatial Genetic Software (SGS, version 1.0d), to detect the spatial genetic autocorrelation using Fst statistics as a genetic distance measure. The construction of genetic distograms has two advantages.[16] First, it describes spatial patterns for multiple variables simultaneously. Second, it applies established concepts of genetic distance to measure dissimilarities. Another advantage of genetic distograms in SGS software is that the statistical significance of spatial genetic autocorrelation can be tested by a permutation procedure.

Acknowledgements

The data collection was supported by NSFC and STCSM to Fudan and a NSF grant to LJ and RD.

References

1 Jin L, Su B: Natives or immigrants: origin and migrations of modern humans in East Asia. *Nat Rev Genet* 2000; **1**: 126–133.
2 Chen R, Ye G, Geng Z *et al*: Revelations of the origin of Chinese nation from clustering analysis and frequency distribution of HLA polymorphism in major minority nationalities in Mainland China. *Yi Chuan Xue Bao* 1993; **205**: 389–398. (in Chinese).
3 Du R, Xiao CJ, Cavalli-Sforza LL: Genetic distances between Chinese groups calculated on gene frequencies of 38 loci. *Sci China C Life Sci* 1998; **28**: 83–89.
4 Cavalli-Sforza LL, Menozzi P, Piazza A: *The History and Geography of Human Genes*. Princeton: Princetion University Press, 1994.
5 Xiao CJ, Du RF, Cavalli-Sforza LL, Minch E: Principal component analysis of gene frequencies of Chinese populations. *Sci China C Life Sci* 2000; **43**: 472–481.
6 Chu JY, Huang W, Kuang SQ *et al*: Genetic relationship of populations in China. *Proc Natl Acad Sci USA* 1998; **95**: 11763–11768.
7 Wen B, Li H, Lu D *et al*: Genetic evidence supports demic diffusion of Han culture. *Nature* 2004; **431**: 302–305.
8 Yao YG, Nie L, Harpending H, Fu YX, Yuan ZG, Zhang YP: Genetic relationship of Chinese ethnic populations revealed by mtDNA sequence diversity. *Am J Phys Anthropol* 2002; **118**: 63–76.
9 Yao YG, Kong QP, Bandelt H-J, Kivisild T, Zhang YP: Phylogeographic differentiation of mitochondrial DNA in Han Chinese. *Am J Hum Genet* 2002; **70**: 635–651.
10 Su B, Xiao J, Underhill P *et al*: Y-chromosome evidence for a northward migration of modern humans into Eastern Asia during the last Ice Age. *Am J Hum Genet* 1999; **65**: 1718–1724.
11 Karafet T, Xu L, Du R *et al*: Paternal population history of East Asia: sources, patterns, and microevolutionary processes. *Am J Hum Genet* 2001; **69**: 615–628.
12 Ding YC, Wooding S, Harpending HC *et al*: Population structure and history in East Asia. *Proc Natl Acad Sci USA* 2000; **97**: 14003–14006.
13 Manni F, Guerard E, Heyer E: Geographic patterns of (genetic, morphology, linguistic) variation: how barriers can be detected by 'Monmonier's algorithm'. *Hum Biol* 2004; **76**: 173–190.
14 Sokal RR, Thomson AB: Spatial genetic structure of human populations in Japan. *Hum Biol* 1998; **70**: 1–22.
15 Manni F, Guerard E: *Barrier vs. 2.2. Manual of the User: Population Genetics Team*. Paris: Museum of Mankind (Musee de I'Homme), [Publication distributed by the authors] 2004.
16 Degen B, Petit R, Kremer A: SGS–Spatial Genetic Software: a computer program for analysis of spatial genetic and phenotypic structures of individuals and populations. *J Hered* 2001; **92**: 447–449.
17 Jorde LB, Bamshad M, Rogers AR: Using mitochondrial and nuclear DNA markers to reconstruct human evolution. *Bioessays* 1998; **20**: 126–136.
18 Wen B, Xie X, Gao S *et al*: Analyses of genetic structure of Tibeto-Burman populations reveals sex-biased admixture in southern Tibeto-Burmans. *Am J Hum Genet* 2004; **74**: 856–865.
19 Wen B, Hong S, Ling R *et al*: The origin of Mosuo people as revealed by mtDNA and Y chromosome variation. *Sci China C Life Sci* 2004; **47**: 1–10.

20 Wen B, Li H, Gao S *et al*: Genetic structure of Hmong-Mien speaking populations in East Asia as revealed by mtDNA lineages. *Mol Biol Evol* 2005; **22**: 725–734.
21 Kivissild T, Tolk H-V, Parik J *et al*: The emerging limbs and twigs of the East Asian mtDNA tree. *Mol Biol Evol* 2002; **19**: 1737–1751.
22 Barbujani G: Geographic patterns: how to identify them and why. *Hum Biol* 2000; **72**: 133–153.
23 Jenks, George F: The data model concept in statistical mapping. *International Yearbook of Cartography* 1967; **7**: 186–190.
24 Monmonier MS: Maximum-difference barriers: an alternative numerical regionalization method. *Geogr Anal* 1973; **3**: 245–261.
25 Cliff AD, Ord JK: *Spatial Autocorrelation*. London: Pion Limited, 1973.
26 Sokal RR, Oden NL: Spatial autocorrelation in biology. 1. Methodology. *Biol J Linnean Soc* 1978; **10**: 199–228.
27 Ned L: *CrimeStat III: A Spatial Statistics Program for the Analysis of Crime Incident Locations (version 3.0)*. Houston, TX: Ned Levine & Associates/Washington, DC, USA: National Institute of Justice, 2004.
28 Su B, Xiao C, Deka R *et al*: Y chromosome haplotypes reveal prehistorical migrations to the Himalayas. *Hum Genet* 2000; **107**: 582–590.
29 Du R, Yip VF: *Ethnic Groups in China*. Beijing: Science Press, 1993. (in Chinese).
30 You Z: *History of Yunnan Nationalities*. Kunming: Yunnan University Press, 1994. (in Chinese).
31 Wang ZH: *History of Nationalities in China*. Beijing: China Social Science Press, 1994. (in Chinese).
32 Ge JX, Wu SD, Chao SJ: *Zhongguo Yimin Shi (The Migration History of China)*. Fuzhou: Fujian People's Publishing House, 1997. (in Chinese).
33 Li SL, Yamamoto T, Yoshimoto T *et al*: Phylogenetic relationship of the populations within and around Japan using 105 short tandem repeat polymorphic loci. *Hum Genet* 2006; **118**: 695–707.
34 Hoffmann MH, Glass AS, Tomiuk J, Schmuths H, Fritsch RM, Bachmann K: Analysis of molecular data of Arabidopsis thaliana (L.) Heynh. (Brassicaceae) with Geographical Information Systems (GIS). *Mol Ecol* 2003; **12**: 1007–1019.
35 Dupanloup I, Schneider S, Excoffier L: A simulated annealing approach to define the genetic structure of populations. *Mol Ecol* 2002; **11**: 2571–2581.
36 Degen B, Scholz F: Spatial genetic differentiation among populations of European beech (Fagus sylvatica L.) in Western Germany as identified by geostatistical analysis. *Forest Genet* 1998; **5**: 191–199.

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