Genetic Structure of Hmong-Mien Speaking Populations in East Asia as Revealed by mtDNA Lineages

Bo Wen, Hui Li,* Song Gao,* Xianyun Mao,* Yang Gao,* Feng Li,* Feng Zhang,* Yungang He,* Yongli Dong,† Youjun Zhang,‡ Wenju Huang,‡ Jianzhong Jin,* Chunjie Xiao,† Daru Lu,* Ranajit Chakraborty,§ Bing Su,§ Ranjan Deka,§ and Li Jin*§*

*State Key Laboratory of Genetic Engineering and Center for Anthropological Studies, School of Life Sciences, Fudan University, Shanghai, China; †Department of Biology and Human Genetics Center, Yunnan University, Kunming, China; ‡Guanxi University of Nationalities, Nanning, China; §Center for Genome Information, Department of Environmental Health, University of Cincinnati, Ohio

Hmong-Mien (H-M) is a major language family in East Asia, and its speakers distribute primarily in southern China and Southeast Asia. To date, genetic studies on H-M speaking populations are virtually absent in the literature. In this report, we present the results of an analysis of genetic variations in the mitochondrial DNA (mtDNA) hypervariable segment 1 (HVS1) region and diagnostic variants in the coding regions in 537 individuals sampled from 17 H-M populations across East Asia. The analysis showed that the haplogroups that are predominant in southern East Asia, including B, R9, N9a, and M7, account for 63% (ranging from 45% to 90%) of mtDNAs in H-M populations. Furthermore, analysis of molecular variance (AMOVA), phylogenetic tree analysis, and principal component (PC) analysis demonstrate closer relatedness between H-M and other southern East Asians, suggesting a general southern origin of maternal lineages in the H-M populations. The estimated ages of the mtDNA lineages that are specific to H-M coincide with those based on archeological cultures that have been associated with H-M. Analysis of genetic distance and phylogenetic tree indicated some extent of difference between the Hmong and the Mien populations. Together with the higher frequency of north-dominating lineages observed in the Hmong people, our results indicate that the Hmong populations had experienced more contact with the northern East Asians, a finding consistent with historical evidence. Moreover, our data defined some new (sub-)haplogroups (A6, B4e, B4f, C5, F1a1, F1a1a, and R9c), which will direct further efforts to improve the phylogeny of East Asian mtDNAs.

Introduction

The languages spoken in East Asia belong to six major linguistic families: Sino-Tibetan (S-T, including Han and Tibeto-Burman subfamilies), Altai, Daic, Hmong-Mien (H-M), Austro-Asiatic (A-A), and Austronesian (Ethnologue 14th Edition: http://www.ethnologue.com/). To a large extent, the linguistic classification corresponds to the historical records regarding the ancestry of East Asians. For example, it has been suggested that populations speaking the Tibeto-Burman (T-B), Daic, H-M, and A-A languages were derived from the ancient tribes *Di-Qiang, Bai-Yue, Nan-Man, and Bai-Pu,* respectively (Wang 1994). Therefore, systematic sampling and analysis of genetic variations of the populations from different linguistic groups may shed light on the genetic structure and population history of East Asians (Su et al. 2000; Wen et al. 2004*a*, 2004*b*).

Hmong-Mien is one of the major language families spoken in southern China and Southeast Asia. It contains 32 languages, and their speakers exceed 12 million in China alone (2000 Census). Archeological and historical studies have shown that proto–H-M populations were associated with the Neolithic cultures that were found in the Middle Reach of the Yangtze River, i.e., *Daxi Culture* (5,300–6,400 years before present [YBP]) and *Qujialing Culture* (4,600–5,000 YBP), and the *San-Miao* tribes in Central-southern China (Fei 1999). However, no systematic genetic study has been reported to reveal the genetic structure and prehistory of the H-M populations.

Key words: East Asia, Hmong-Mien, Genetic structure, mtDNA, haplogroup.

E-mail: lijin@fudan.edu.cn; li.jin@uc.edu.

Mol. Biol. Evol. 22(3):725–734. 2005
doi:10.1093/molbev/msi055
Advance Access publication November 17, 2004

Genetic differentiation between the southern and northern East Asian populations has been observed with classic markers (Cavalli-Sforza, Menozzi, and Piazza 1994; Xiao et al. 2000), mitochondrial DNA (mtDNA) (Yao et al. 2002*a*; Kivisild et al. 2002), and Y chromosome variations (Su et al. 1999; Jin and Su 2000; Karafet et al. 2001), although the mechanism that led to such differences remains controversial (Su et al. 1999; Ding et al. 2000; Jin and Su 2000; Karafet et al. 2001). The final revelation of the mechanism of the north-south division in East Asian peoples warrants an extensive accumulation of genetic data from East Asian populations, particularly from the understudied groups, such as the H-M-speaking populations.

Haploid and maternally inherited mtDNA is one of the most powerful tools in reconstructing the evolutionary history of human populations (Wallace, Brown, and Lott 1999; Cavalli-Sforza and Feldman 2003). The major structure of the phylogeny of mtDNA haplotypes found in East Asia is becoming increasingly robust, owing to the effort of combining the variations in the control and coding regions (Wallace, Brown, and Lott 1999; Yao et al. 2002*a*; Kivisild et al. 2002) and availability of increasing numbers of complete mtDNA sequences (Kong et al. 2003*a*). Specifically, macro-haplogroup M and N encompass almost all mtDNA lineages in East Asia. M encompasses the lineages of haplogroup D, G, M7, M8 (ancestral to the haplogroup C and Z), M9 etc., whereas N is composed of the lineages of A, B, R9 (ancestral to haplogroup F), and N9. By studying the distribution of these haplogroups and their sub-haplogroups, one can trace the migration of populations in East Asia, as exemplified by a recent account of peopling of Japan and Korea (Kivisild et al. 2002).

In this report, we present the results of a study of mtDNA diversity of the variations in the hypervariable

Table 1
H-M Populations in This Study

Population [a]	*n*	Ethnicity [b]	Location	Language
Hmong				
1. Miao-Hunan (MHN)	103	Miao	Jishou, Hunan	HMONG, Xiangxi
2. Miao-Yunnan (MYN)	39	Miao	Wenshan, Yunnan	HMONG, Chuanqiandian
3. Bu Nu (YBN)	19	Yao	Dahua, Guangxi	BU-NAO, Bunu
4. Mu Bin (YMB)	6	Yao	Tianlin, Guangxi	BU-NAO, Bunu
Mien				
5. Ba Pai (YBP)	35	Yao	Liannan, Guangdong	BAI PAI, Zaomin
6. Din Ban (YDB)	10	Yao	Mengla, Yunnan	IU MIEN, Mian-Jin
7. Guo Shan (YGS)	24	Yao	Jianghua, Hunan	IU MIEN, Mian-Jin
8. Hua Tou (YHT)	19	Yao	Fangcheng, Guangxi	IU MIEN, Mian-Jin
9. Mien (YMI)	32	Yao	Shangsi, Guangxi	IU MIEN, Mian-Jin
10. Pan (YPA)	32	Yao	Tianlin, Guangxi	IU MIEN, Mian-Jin
11. Tu Yao (YTU)	41	Yao	Hezhou, Guangxi	IU MIEN, Mian-Jin
12. Xi Ban (YXB)	11	Yao	Fangcheng, Guangxi	IU MIEN, Mian-Jin
13. You Mian (YYM)	27	Yao	Mengla, Yunnan	IU MIEN, Mian-Jin
14. Kim Mun (YKM)	40	Yao	Malipo, Yunnan	KIM MUN, Mian-Jin
15. Lan Tin (YLT)	26	Yao	Tianlin, Guangxi	KIM MUN, Mian-Jin
16. Lowland (YLO)	42	Yao	Fuchuang, Guangxi	KIM MUN, Mian-Jin
17. Wuzhou (YWU)	31	Yao	Fuchuang, Guangxi	KIM MUN, Mian-Jin

[a] The abbreviated name of each population is given in the parentheses.
[b] The official ethnicity.

segment 1 (HVS1) region, along with 16 diagnostic variants in the coding regions, in 537 individuals sampled from 17 H-M-speaking populations across East Asia. This is the first systematic study of the genetic structure of the H-M populations, aimed to investigate the matrilineal ancestry of the H-M populations and to provide more data for evaluating the genetic structure of East Asians. The understanding of the distribution of mtDNA variations also bears importance in the studies of mtDNA-associated diseases (Wallace, Brown, and Lott 1999).

Material and Methods

Samples

Blood samples of 537 unrelated individuals from 17 H-M populations were collected with appropriate informed consent. These populations encompass two major linguistic branches (subfamilies), i.e., Hmong and Mien, with diverse geographic distribution in Guangxi, Guangdong, Hunan, and Yunnan provinces, China. The Hmong-speaking and Mien-speaking populations are often referred to as Miao and Yao, respectively. Note that discrepancy exists between linguistic classification and official ethnicity. For example, two Hmong-speaking populations (Bu Nu and Mu Bin) are classified as Yao. The detailed sample information including the names of populations, sample size, official ethnicity, geographic location, and linguistic branches is given in table 1 and figure 1.

Genotyping of mtDNA Polymorphisms

Genomic DNA was extracted by standard phenol/chloroform methods. A fragment containing the HVS1 region was amplified by primers L15974 and H16488 (Yao et al. 2002*a*), and the purified PCR product was sequenced by BigDye terminator cycle sequencing kit and ABI 3100 genetic analyzer (Applied Biosystem, Foster City, Calif.), and the same primers were used for amplification. Most of samples were sequenced in both directions to confirm the status and position of the variants. Primers were also designed for amplifying multiple fragments containing 16 polymorphisms in the coding regions, and the PCR products were digested by restriction enzymes: 663 *Hae*III, 3394 *Hae*III, 5176 *Alu*I, 4831 *Hha*I, 7598 *Hha*I, 9824 *Hinf*I,13259 *Hinc*II, 10394 *Dde*I, 10397 *Alu*I, and 12406 *Hpa*I, except for the 9bp deletion, which was genotyped by electrophoresis of the polymerase chain reaction (PCR) product (Wallace, Brown, and Lott 1999; Yao et al. 2002*a*; Kivisild et al. 2002). The additional variations, at positions 3010, 4715, 5417, 12705, and 10310 were genotyped by sequencing or PCR–restriction fragment length polymorphism (RFLP) assay by engineering restriction sites in the primers (primer information and genotyping protocol for all the coding region variants are available on request).

Data Analysis

The 360-bp-long HVS1 sequences (16024–16383) were edited and aligned against the revised Cambridge Reference Sequence (CRS; Andrews et al. 1999) using DNASTAR software (DNASTAR, Inc.). For quality control, we rechecked all the polymorphic sites in the chromatogram using DNSTAR software. And then, in comparison with the established phylogeny (Kivisild et al. 2002), the data set was checked using NETWORK software (www.fluxus-engineering.com) to test for potential errors (Bandelt et al. 2001). Overall, 537 HVS1 sequences from 17 H-M populations have been submitted to GenBank (Accession numbers AY546723–AY547259).

Haplotype diversity (H) and nucleotide diversity (π) were used to evaluate the extent of within-population variation (Nei 1987). Probability of identity (Nei 1987) and F_{ST} were calculated to investigate between-population diversity. The genetic structure of populations was investigated by the analysis-of-molecular-variance approach (AMOVA; Excoffier, Smouse, and Quattro 1992). The calculations of diversity indices, F_{ST} and AMOVA were

conducted on ARLEQUIN software (Schneider et al. 2000). An unrooted Neighbor-Joining tree was constructed by F_{ST} distances using MEGA 2.1 software (Kumar et al. 2001). The data sets of other East Asian populations—4 A-A (Yao and Zhang 2002; L.J., unpublished data), 6 Altai (Yao et al. 2002*b*; Kong et al. 2003*b*), 10 Daic (Yao et al. 2002*b*; L.J., unpublished data), 10 northern Han, 15 southern Han (Oota et al. 2002; Yao et al. 2002*a*; Wen et al. 2004*b*), and 17 T-B populations (Yao et al. 2002*b*; Wen et al. 2004*a*)—were integrated in these analyses. Correlation of genetic and geographic distance was tested by Mantel test, using ARLEQUIN software. Significance of correlation coefficient (r) was obtained by comparing with distribution of 1,000 random permutations.

Haplogroup affiliation of each mtDNA sequence was inferred based on the HVS1 motif and diagnostic variants in the coding regions, following Kivisild et al. (2002) and Kong et al. (2003*a*). Median joining networks (Bandelt, Forster, and Rohl 1999) were constructed by NETWORK software to investigate detailed relationships of the lineages within each haplogroup and to infer possible new haplogroups wherever applicable. The nomenclature of newly established (sub)-haplogroups followed Kong et al. (2003*a*). Coalescence time and its standard error of the haplogroups were calculated by the methods developed by Forster et al. (1996) and Saillard et al. (2000), respectively. To show genetic relationships between H-M and other East Asian populations, principal component analysis (PCA) was conducted by haplogroup frequencies, using SPSS10.0 software [SPSS Inc., Chicago, Ill.]). To avoid bias from estimating of haplogroup frequencies based on HVS1 sequences alone (Kivisild et al. 2002), in the PCA analysis, we used only the data sets typed in both the HVS1 and the coding regions (Yao et al. 2002*a*; 2003; Kivisild et al. 2002; Kong et al. 2003*b*; Wen et al. 2004*a*, 2004*b*; this study; L.J., unpublished data).

FIG. 1.—Map of China and Southeast Asia showing geographic locations of the H-M populations sampled. The numbers of populations are consistent with table 1.

Results

Population Diversity and Haplotype Sharing in H-Ms

In the 537 samples studied, 271 HVS1 haplotypes were observed, 78 of which are shared by more than two individuals. The haplotype diversity (H) and nucleotide diversity (π) estimates are presented in table 2. In the H-M populations, H ranges from 0.928 to 1, and π from 0.017 to 0.023, diversity ranges similar to those observed in other ethnic groups in East Asia (Yao et al. 2002*b*). Some populations such as MYN, YMI, YGS, and YTU have very low haplotype diversity, indicating possible influence of drift resulting from bottleneck events.

Of the 271 HVS1 haplotypes observed in H-M, 61 are shared by more than two populations. As shown in table 2,

Table 2
Diversity Indices of H-M Populations

Population	Size	Hap [a]	Unique Hap [b]	H [c]	π [d]
Hmong					
MHN	103	85	59 (69%)	.995 ± .002	19.75 ± 10.34
MYN	39	22	10 (45%)	.937 ± .022	17.79 ± 9.56
YBN	19	15	2 (13%)	.977 ± .023	17.41 ± 9.66
YMB	6	6	1 (17%)	1.000 ± .096	21.85 ± 13.69
Mien					
YBP	35	24	13 (54)	.975 ± .013	21.46 ± 11.38
YDB	10	8	8 (100%)	.956 ± .059	20.00 ± 11.58
YGS	24	15	7 (47%)	.928 ± .039	17.01 ± 9.34
YHT	19	14	5 (36%)	.965 ± .028	23.46 ± 12.69
YMI	32	17	5 (29%)	.954 ± .017	17.60 ± 9.53
YPA	32	27	10 (37%)	.988 ± .012	21.87 ± 11.62
YTU	41	18	4 (22%)	.944 ± .015	18.01 ± 9.66
YXB	11	11	8 (73%)	1.000 ± .039	21.11 ± 12.04
YYM	27	19	12 (63%)	.963 ± .021	19.92 ± 10.73
YKM	40	33	14 (42%)	.990 ± .008	17.83 ± 9.58
YLT	26	19	7 (37%)	.966 ± .022	20.98 ± 12.74
YLO	42	37	26 (70%)	.993 ± .007	21.40 ± 11.30
YWU	31	29	19 (66%)	.994 ± .011	20.25 ± 10.84
All samples	537	271		.990 ± .001	19.89 ± 10.33

NOTE.—Underlined items indicate populations with low halotype diversity.

[a] Number of haplotypes.

[b] Number of unique haplotypes.

[c] Haplotype diversity.

[d] Nucleotide diversity multiplied by 1,000.

Table 3
Haplotype Matching Probabilities Between H and Ms

	1	2	3	5	6	7	8	9	10	11	12	13	14	15	16
1. MHN															
2. MYN	1.4														
3. YBN	.4	2.2													
5. YBP	.5	.9	.9												
6. YDB	0	0	0	0											
7. YGS	.7	5.3	1.7	1.1	0										
8. YHT	.8	1.1	1.4	1.4	0	1.3									
9. YMI	.8	1.8	2.6	1.3	0	2	2.1								
10. YPA	.4	1.3	1.8	1	0	2.1	1.6	1.2							
11. YTU	.6	3.3	.8	.8	0	4.2	1	1.1	1						
12. YXB	0	0	0	0	0	0	1	.9	.3	0					
13. YYM	.5	.8	.7	.4	0	.7	1.2	.9	.7	1.2	0				
14. YKM	.4	2.3	2.1	0	0	3	.6	2.7	1	1.5	0	.5			
15. YLT	.1	1	1.4	.7	0	1.5	2.2	2.9	1.3	.8	1.7	.5	1.4		
16. YLO	.4	1.2	.5	.2	0	1.7	.4	.6	.4	.7	0	.2	.8	.3	
17. YWU	.4	.6	1	.4	0	.5	.7	.9	1.2	.6	0	.5	.7	.8	.1

NOTE.—All values have been magnified 100 times. YMB (Table 1, number 4) was not included in these analyses because of its small sample size ($n = 6$).

numbers of population-specific haplotypes are generally low in the H-M populations. The proportion of population-specific haplotypes ranges from 13% (YBN) to 72% (YXB), with 10 of 17 populations below 50%. We further investigated this pattern by calculating probability of identity (M), which is the probability that two random sequences, one each from the two populations, are identical, without adjusting for within-population probability of identity (table 3). Almost all population pairs have the M values higher than zero, except for those involving YDB and YXB, whose sample sizes are relatively small. Note that M's among H-M populations are higher than those between other East Asian populations reported by Yao et al. (2002*b*).

Possible New Haplogroups Implied by H-M mtDNA datD

The mtDNA genealogies observed in this study are consistent with the major phylogenetic structure of the East Asian mtDNAs, defined by Kivisild et al. (2002), and modified by Kong et al. (2003*a*). However, the results in this study offer an opportunity for further refinement of the structure with the observation of some new haplotypes (the HVS1 motif and genotyping result of the 16 variations in coding regions are given in the Supplementary Material online). In the following listing, we provide our suggestion of these new haplogroups, following the nomenclature of Kong et al. (2003*a*): (1) four R9 H-M mtDNAs carry the 16157–16304 motif without the 10310 transition. This motif was also observed in some populations in Taiwan (Tajima et al. 2003), southwestern China (Yao et al. 2002*a*), and Southeast Asia (Oota et al. 2002). We define this lineage as haplogroup R9c. (2) 38% (6 of 16) of the haplogroup A mtDNAs bear the 16126–16223–16290–16319 motif, which is also found in the Chinese Han (Horai et al. 1996; Oota et al. 2002), Korean (Lee et al. 1997), and Japanese (Horai et al. 1996) with very low frequencies. We refer it as haplogroup A6. (3) The motif 16147–16184A–16189–16217–16235 was found in 10 H-M B4 mtDNAs, and is almost absent in other populations except in one sample in the Dai (Yao et al. 2002*b*) and one in the Vietnamese (Oota et al. 2002). It is likely to be a new B4 haplogroup, which will be referred to here as B4e. (4) Four B4 mtDNAs carry the16140–16189–16217–16274 motif, which is relatively frequent in southern East Asians, especially in Taiwan aborigines (Tajima et al. 2003). We name it B4f. (5) Some 40% of haplogroup C H-M samples have the motif 16189–16298–16327, which will be referred to as C5. It is almost completely absent in the northern East Asian populations and seems to be the major branch of haplogroup C in the southern East Asians such as Dai, Zhuang, and Lahu (Yao et al. 2002*b*; Yao and Zhang 2002). (6) Most of (29/45) F1a H-M samples carry the motif 16162–16172–16304, and nearly half of these (14/29) bear the 16108 transition at the same time. They are named as F1a1 and F1a1a, respectively. These two lineages are extensively distributed in East Asia and account for about 50% of total F1a mtDNAs. We believe that more variations in the coding regions need to be studied for further characterizing the phylogenetic positions of these newly defined haplogroups, although all of them appear to be mono-phylogenetic in the median-joining networks (data not shown).

Of the 537 mtDNAs found in H-M, 9.5% (42 M*, 1 N*, and 11 R*) cannot be assigned to the presently defined haplogroups. Three M* mtDNAs share the HVS1 haplotype 16111–16223–16235–16362, a motif that is also found in two Zhuang samples (Yao et al. 2002*b*). The motif 16192–16223–16292, an analog of west Euroasian haplogroup W, is observed in four M* samples, but it has not been seen in other East Asia populations. Ten Miao M* samples share the motif 16093–16223–16311–16362–16381. This haplotype is very rare in other populations except for Naxi (Wen et al. 2004*a*), where its frequency is considerable (9%). Four R* mtDNAs carry both the 16189–16311 motif and 10394 DdeI+. More effort is required to characterize these haplotypes.

Distribution of Haplogroups in the H-M Populations

Based on the phylogeny of East Asian mtDNA lineages (Kivisild et al. 2002; Kong et al. 2003*a*; this

Table 4
Haplogroup Distribution Among H-M Populations

	Population [a]																
Haplo-group	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
A	7						1									1	1
A6	2				1					1	1			1			
B					1												
B4	2				1	3	2				6					2	
B4a	5	5	5		1	1		2	5	2	3	1	7	3	3	3	4
B4b1	3				2			1				1	5	1	1	3	1
B4e	1				3	1		1	2					2			
B4f	1					1				1					1		
B5														1			
B5a	6	8	1	1	7		7	3	3	4	7		1	5	2	1	3
B5b	1									1						1	
C	8	2						2						2		2	3
C5	2								4			1		2	3		1
D	15	6	2		2		4	1	5	1	3		7	6	1	7	1
D5	3	1					1						1		1		
D5a	2			1												1	
F	3		1					1				1	1		1		
F1																1	
F1a	2				1	1	1	1		1	1	1			5	2	
F1a1	4			1				1	2	2				1	1	2	1
F1a1a	3		1		1			2	1	2	4						
F1b	2	1	1	1	3		1			2		1				1	
F1c										1						1	
F2a																4	
F3	1	1	1		2		1			2						1	1
G	2	1								1							
G2	1							2	2		4	1		2	1		
G3	2																
M*	7	9	1		3		4	1		1	3	1	1	7		4	
M7	1	1	2		2				1		3		1				1
M7b	4	1	2						4		2	1	1		1	2	3
M7b1			1						1	4	2	1	1	1	1		5
M7c1			1	1													
M8																	1
M8a	2				1												
N*	1																
N9a	2	1		1	2	2	2			1						2	2
R*	1				2	1				2		1			2		2
R9										1				2		1	
R9b	4	2						1	2	1	2		1	1	2		
R9c	1													2			1
Y	1																
Z	1									1				1			

[a] The population IDs are consistent with table 1.

study), we inferred haplogroup affiliation for every mtDNA in this study, by considering both HVS1 and coding region variations. The haplotype frequencies in the H-M populations are presented in table 4. Some 43% of H-M mtDNAs belong to haplogroup M and its derived haplogroups (M7, M8, M8a, C, Z, D, and G); and the macro-haplogroup N and its nested lineages (A, B, F, R9, R, N9a, Y) account for 57% H-M mtDNAs.

Haplogroup D, M7, and M8 are the major M types in H-M, which account for 13.4%, 9.7%, and 7.3% of the total samples, respectively. Haplogroup D was found in most of the H-M populations, except for YDB and YXB. The majority (78%) of the H-M haplogroup D mtDNAs belong to D4 branch. It is surprising that 23 D4 mtDNAs from 10 H-M populations share the same motif 19092–16223–16362, which is very rare in other populations (only 1 Taiwan Han and 1 Zhuang). Coalescence time of the D4 mtDNAs with this motif is estimated at 4,000 ± 2, 000 YBP. M7 is also present in most H-M populations, with its frequency ranging from 2.5% to 32%. M7b (including M7b1) accounts for 73.8% of M7 mtDNAs, making it the most common M7 mtDNA type in H-M. In contrast, the distribution of M8 is much more restricted, and 35% (6/17) of the H-M populations do not have M8. C is the most prevalent M8 type in the H-M populations (82%), whereas the M8a and Z are much rarer. Haplogroup G is sporadically distributed in the H-M populations with low frequency (3.5% in total), of which G2 is the most common type (68%). Interestingly, all the H-M G2 samples also bear the coding region variation 7598 Hha−, which was previously considered as the diagnostic mutation of the haplogroup E. The co-occurrences of 4831 Hha+ and 7598 Hha− were also observed in the Korean G2 samples (Snall et al. 2002).

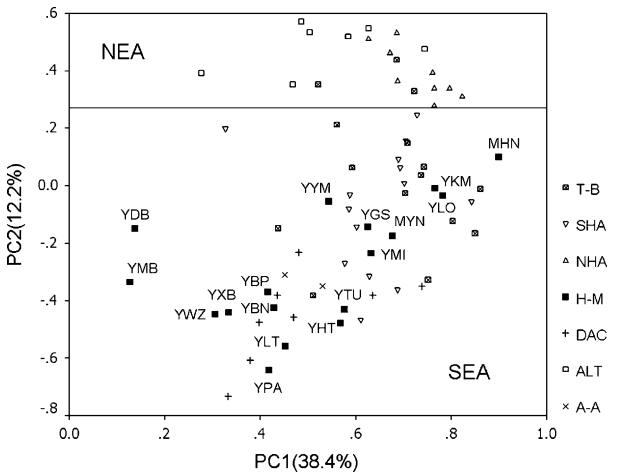

FIG. 2.—PC plot of mtDNA haplogroup frequencies. Abbreviations for populations: T-B: Tibeto-Burman; H-M: Hmong-Mien; SHA: Southern Han; NHA: Northern Han; DAC: Daic; ALT: Altai; A-A: Austro-Asiatic.

In the macro-haplogroup N, B and R9 (including F, R9b, and R9c) are the predominant lineages, with very high frequencies (30.2% for B and 19.8% for R9). Frequency of haplogroup B in the H-M populations is the highest in mainland East Asians (Yao, Watkins, and Zhang 2000). All the H-M populations carry haplogroup B mtDNAs, with its frequency ranging from 17% (in YMB) to 60% (in YDB). The B4 (18.2%) and B5 (11.7%) haplogroups encompass almost all haplogroup B mtDNAs in H-M, and B4a (9.3%) and B5a (11%) account for one third of B mtDNAs. B5a is very homogeneous in H-M: 42 of 59 B5a mtDNAs share the motif 16140–16189–16266A, and the MJ network of B5a indicates a typical star-like phylogeny, whose coalescence time is estimated at 6,000 ± 2, 000 YBP. B4a itself contains many branches, which need to be further investigated. Over one third (38%, i.e., 19/50) of B4a mtDNAs in H-M bear the motif 16129–16189–16217–16261, which is rare in the other East Asian populations (only seen in 2 Shandong Han, 1 Yunnan Han, 1 Dai, and 1 Thai). Thus the time estimation of the age this lineage may contribute to understanding the peopling of the H-M populations. The time estimation for this lineage is 3,560 ± 2,050 YBP. The motif 16189–16217–16261–16357 is completely absent in other populations, and thus it turns out to be a H-M–specific lineage. It was found in 7 H-M mtDNAs, and the time estimation is 26,000 ± 16,000 YBP. It should be noted that this estimate is somewhat imprecise because of the small sample size. B4e is another H-M–specific lineage, which was found in 10 H-M mtDNAs and almost absent in the other populations (seen only in 1 Dai and 1 Vietnamese). Its age is 13,500 ± 9,500 YBP. Other haplogroup B lineages, B4b1, B4f, B5b, and B5b are distributed sporadically, with low frequencies in H-M.

Haplogroup F was observed in all the H-M populations (15.5% in total), and it appears to be the most frequent haplogroup in R9. More than half (54%) of F mtDNAs belong to the F1a lineage, which is the predominant F type in H-M and other southern East Asians (Kivisild et al. 2002). F1a1 and F1a1a account for one-third of F1a samples respectively. The estimated ages for H-M F1a1 and F1a1a are 18,000 ± 10,400 YBP and 8,600 ± 6,100 YBP, respectively. R9c, a newly defined haplogroup in this study, is observed only in 4 H-M samples. This rare haplogroup is found only in southern China and Southeast Asia. By including an additional 11 R9c samples (Oota et al. 2002; Yao et al. 2002*a*; Tajima et al. 2003; B.W, unpublished data), the age of R9c is estimated as 29,600 ± 16,300 YBP, appearing to be a deep lineage distributed in southern East Asia. Other R9 lineages, F1b, F1c, F2a, and F3 are seen in some populations with low frequencies.

N9 (including N9a and Y) and A are the other two haplogroups under the N branch. Haplogroup A was observed in 3% (16 of 537) of H-M mtDNAs, 56% (9/16) of which are concentrated in MHN. Frequency of haplogroup A in MHN is very high (8.7%), which is close to the frequency in northern populations. In contrast, it is absent or occurs as singletons in the other H-M populations. Haplogroup A6, which was observed in the Northern Han, Japanese, and Korean populations, occurs in some H-M populations as singletons. N9 consists of Y and N9a. Only one sample in the MHN belongs to Y, and N9a occurs in some populations with low frequency.

Population Cluster as Revealed by PCA

Figure 2 presents the PCA results conducted in H-M and other East Asian populations. Northern East Asians (NEA) and southern East Asians (SEA) are clearly separated by PC2 (accounting for 12.2% of the total variation), and the H-M populations fall entirely into the southern group. The H-M populations clustered together with other SEA groups (Daic, A-A, Southern Han [S. Han], and Southern Tibeto-Burman [STB]), indicating a general resemblance of the maternal lineages between H-Ms and other SEAs. Some H-Ms (MHN, MYN, YKM, YLO, YYM, YGS, and YMI) are closer to S. Han and STB, whereas the rest clustered with Daic and Austro-Asiatic groups, two typical SEA groups. Interestingly, Miao-Hunan (MHN) is closer to NEAs than other H-M populations. Its PC2 value (0.10) is higher than all the other H-Ms (PC2, mean 0.31, range −0.01, −0.64), although the difference is only marginally significant (*t*-test, $P = 0.061$). To explore possible influence of the difference of sample size on the PC results, we (1) removed those populations whose sample size was <20 in the PC analysis; (2) resampled 30 individuals from MHN ($n = 103$) and conducted PCA 20 times. These gave almost the same PCA results, indicating sample size has little influence on the PCA.

Between-Population Diversity as Revealed by HVS1 Sequences

The average F_{ST}s of H-M versus other East Asian populations were estimated to examine their genetic relationships (table 5). The average F_{ST}s of H-M and NEA populations are significantly larger than those between H-M and SEAs (*t*-test, $P < 10^{-3}$). Almost all H-M

Table 5
Average Fst of H-Ms versus Other East Asian Populations

	F_{ST} with NEAs			F_{ST} with SEAs					
	N.Han	Altai	NEA	A-A	Daic	S. Han	STB	SEA [a]	SEA [b]
MHN	.009	.017	.019	.041	.019	.009	.021	.019	.025
MYN	.045	.055	.056	.096	.048	.040	.059	.053	.061
YBN	.043	.059	.061	.048	.005	.014	.044	.025	.016
YBP	.075	.087	.091	.093	.037	.044	.070	.056	.052
YDB	.110	.124	.130	.128	.059	.062	.101	.081	.077
TGS	.049	.061	.063	.090	.034	.033	.057	.047	.049
YHT	.052	.072	.072	.037	.014	.021	.043	.029	.020
YMI	.035	.047	.051	.053	.007	.012	.039	.024	.019
YPA	.047	.070	.067	.024	.008	.021	.041	.025	.012
YTU	.061	.079	.079	.052	.016	.028	.054	.037	.026
YXB	.048	.069	.069	.057	.007	.016	.041	.027	.021
YYM	.106	.118	.122	.134	.067	.068	.103	.086	.085
YKM	.024	.033	.036	.052	.013	.012	.034	.024	.023
YLT	.076	.096	.097	.045	.016	.034	.060	.040	.024
YLO	.021	.034	.035	.034	.012	.010	.025	.018	.018
YWZ	.045	.054	.060	.056	.007	.017	.051	.030	.020
Average	.057	.072	.074	.064	.022	.028	.055	.039	.033

[a] Including southern Han (S. Han) and STB.
[b] Excluding southern.Han and STB.

populations have a lower F_{ST} with SEAs than that with NEAs, except for MHN and MYN, which show very similar F_{ST}s with the two groups. Of the 16 H-M populations analyzed, MHN shows the lowest F_{ST} with NEAs, suggesting its closer affinity with NEAs. The differentiation between H-M and other linguistic groups was also investigated. For two NEA groups, i.e., northern Han and Altais, the former is closer to H-Ms than the latter (*t*-test, $P < 10^{-3}$). For the SEA groups, H-Ms are much closer to Daic and S. Han than they are to STB and A-A (*t*-test, $P < 10^{-3}$ for all comparisons).

An unrooted NJ tree was constructed among East Asian populations using F_{ST} distances. Because the STBs and S. Hans have experienced markedly north-south admixtures in their maternal lineages during the past 2,500 years (Wen et al. 2004*a*, 2004*b*), we excluded them from the phylogenetic analysis. In figure 3, NEAs (Altai, Northern Han, and Northern T-B) and SEAs (A-A, Daic) are largely divided into two different clusters, and almost all H-M populations grouped with the SEA cluster. All the A-A populations formed one single branch, departing from H-Ms and Daics. It was noted that MHN lies between NEAs and SEAs, being closer to SEA populations than to NEAs.

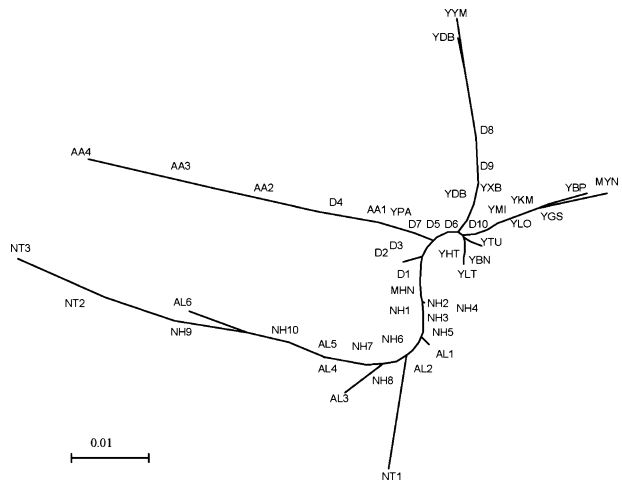

FIG. 3.—Unrooted Neighbor-Joining tree of the H-M and other East Asian populations. Key to prefixes: AA, A-A; AL, Altai, D, Daic; NH, Northern Han; NT, Northern T-B.

The genetic structure between H-Ms and other East Asians were tested by AMOVA (table 6). Significant structures are tested in all grouping strategies, although the vast majority of variances (94.0% to 97.9%) are due to difference within populations. The difference between NEAs and H-Ms is nearly fivefold higher than that between H-Ms and SEAs, suggesting that H-Ms bear much closer affinity with SEAs than with NEAs. We further investigated the structure between different linguistic groups of SEAs. The differences between H-M and Daic (0.49%) and S. Han (0.75%) are much lower than those between STB (1.63%) and A-A (3.86%), demonstrating that the close affinity of H-M with SEAs is mainly due to its similarity to Daic and S. Han.

Table 6
AMOVA Result

	Percentage of Variance (*P* value)		
Grouping	Among Groups	Among Population Within Groups	Within Populations
H-M vs. NEAs	3.69 ($<10^{-5}$)	1.86 ($<10^{-5}$)	94.46 ($<10^{-5}$)
H-M vs. SEAs	.78 (.00033)	2.13 ($<10^{-5}$)	97.09 ($<10^{-5}$)
H-M vs. NEAs vs. SEAs	3.03 ($<10^{-5}$)	1.91 ($<10^{-5}$)	95.06 ($<10^{-5}$)
H-M vs. A-A	3.86 (.00050)	2.11 ($<10^{-5}$)	94.03 ($<10^{-5}$)
H-M vs. Daic	.49 (.00919)	1.61 ($<10^{-5}$)	97.91 ($<10^{-5}$)
H-M vs. S. Han	.75 (.00225)	1.64 ($<10^{-5}$)	97.60 ($<10^{-5}$)
H-M vs. STB	1.63 ($<10^{-5}$)	3.24 ($<10^{-5}$)	95.13 ($<10^{-5}$)

NOTE.—*P* values are obtained by 10,000 permutations.

Genetic Distance versus Geographic Distances

Isolation-by-distance (IBD) is one model accounting for microevolutionary processes among populations. Under pure IBD, the differentiation among populations is only determined by drift and short-range gene flows, and a positive correlation between genetic and geographic distances is expected (Cavalli-Sforza, Menozzi, and Piazza 1994). In the H-M populations, the genetic and geographic distances are significantly correlative ($r = 0.404$, $P = 0.015$), suggesting that genetic variations in those populations were at least partially shaped by IBD: gene flow between adjacent populations is more evident than gene flow between distant ones.

Discussion

In the major East Asian mtDNA haplogroups observed in H-M, haplogroups A, D, G, and M8 are prevalent in the NEAs, whereas B, M7, R9, and N9a are dominating in the SEAs (Kivisild et al. 2002; Yao et al. 2002*a*; Kong et al. 2003*b*). The haplogroups M*, N*, and R* may contain some undefined lineages; their distribution has not been well characterized, and they are therefore referred to as "uncertain" here. The lineages that are prevalent in the SEA represent the majority (63%) of the mtDNA gene pool in the H-M populations, ranging from 45% (MHN) to 90% (YDB). If we remove the "uncertain" lineages, 70% H-M mtDNAs belong to the southern lineages ranging from 49% (MHN) to 100% (YDB). The haplogroups prevalent in the NEA account for only 27% of the H-M mtDNAs, ranging from 0% to 47%. Of note, haplogroup D is also extensively distributed in the SEA with moderate frequency (15%, L.J. unpublished results), although its frequency is higher in the NEA (26%; L.J. unpublished results). More efforts should be made to elucidate the genetic history of this lineage. When haplogroup D is removed from the analysis, the frequency of northern lineages in H-M decreases from 27% to 14%, ranging from 0% to 27%. The proportion of SEA lineages in H-Ms is almost equal to that observed in other SEAs (53%, and 66% when including and excluding STB and S. Han), but it is significantly higher than those observed in NEAs (23%, *t*-test, $P < 10^{-3}$). This observation, together with the close affinities with SEAs revealed by average FST, phylogenetic tree analysis, and PCA, suggests southern origins of H-M populations.

Archeologists and historians have suggested that the proto-H-M might be linked with the Neolithic culture in the Middle Reach of the Yangtze River in southern China (Fei 1999), including the *Daxi Culture* (5,300–6,400 YBP) and the *Qujialing Culture* (4,600–5,000 YBP). Haplogroup B5a, which is very homogeneous and shows a star-like phylogeny, accounts for 11% of H-M mtDNAs and exists in most of the H-M populations. The estimated age of B5a in H-M is 6,000 ± 2, 000 YBP. Another lineage, the motif 16129–16189–16217–16261, is strongly H-M–specific and is present in 8 of the 17 H-M populations. The estimated age of this lineage is 3,560 ± 2, 050 YBP. Furthermore, 23 D4 mtDNAs from 10 H-M populations share the same motif ,19092–16223–16362, which is very rare in other populations. Coalescence time of the D4 mtDNAs with this motif is estimated at 4,000 ± 2, 000 YBP. All these estimations are well in line with the age of the two aforementioned cultures excavated in southern China. The study of ancient DNA from the *Daxi Culture* and *Qujialing Culture* would be of great importance to verify the possible affiliations of H-M and these Neolithic cultures.

We have demonstrated that most H-M maternal lineages are of southern origin. The genetic relationship among H-Ms and other SEAs is another important issue that needs to be investigated. The results of average F_{ST} and AMOVA clearly indicated that H-M is closer to Daic and S. Han than to STB and A-A. Their geographic distributions could offer a reasonable explanation. Daic and S. Han are mainly distributed in Central-southern China, which was also populated by H-Ms, according to historical records. In contrast, A-As and STBs were mainly distributed in Southwestern China, an area to which H-Ms moved just during the past several hundreds years (Cang 1997). The Mantel test further demonstrated that the genetic and geographic distances correlate significantly, reflecting the influence of IBD, that is, that gene flows between H-Ms and their adjacent SEA populations had shaped their maternal genetic landscape to some extent.

The Miao populations are relatively distant from the Yao populations, as revealed by F_{ST} analysis and population tree analysis, as well as by the relatively lower frequencies of southern lineages (44.7% for MHN and 51.3% for MYN). In particular, MHN carries the highest frequency of the northern lineages with or without including haplogroup D (47% including D, 27% not including D). The proportion of northern lineages in MHN is significantly higher than that in other H-Ms (*t*-test, $P = 0.017$). This observation may be biased because the sample size of MHN (103) is much larger than the sample sizes of other H-Ms (average 27). We investigated the influence of the difference of sample sizes by resampling 30 individuals from HMN data 500 times. The resampling showed very similar proportions of NEA mtDNAs in MHN (average 0.464; 95% CI, 0.455–0.472) to the real data (0.467), demonstrating no marked influence of sample size on our observation. Again, comparison of genetic distances, the population cluster based on the phylogenetic tree, and the PCA plot demonstrate a closer affinity of MHN for NEAs.

A famous legend concerning the ancient *San-Miao* tribe, thought to be the ancestor of present-day H-M populations, is of great interest. The *San-Miao* tribe expanded northward to the Yellow River drainage area; then, led by the Chiyou, they battled against the *Yan-Huang* tribe (one of the primary Sino-Tibetan ancestors) in Zhulu (in present-day Hubei Province near Beijing). The Chiyou were defeated and pushed back to the south before the *Yan-Huang* dominated the northern China. Our mtDNA data might provide some clues for tracing this march. As we showed earlier, the southern lineages account for only about 50% of Miao mtDNAs; most of the lineages prevalent in NEA are found in Miao-Hunan (MHN), which has the highest frequency of such haplogroups in the H-M populations. A careful inspection

of the distribution of the northern mtDNA lineages revealed more information. A6 is almost absent in other southern populations, but it is present in Miao-Hunan. C5 is the dominating haplogroup C type in the southern populations; however, almost all haplogroup C mtDNAs are non-C5 in the two Miao populations. G3 is a very rare in NEA, and it is completely absent in the south. Surprisingly, two Miao-Hunan mtDNAs carry this haplogroup. These observations suggest that the Miao (Hmong) people may have more contact with the NEA.

In summary, we demonstrated that southern lineages account for the majority of the H-M mtDNA gene pool, a finding consistent with the southern origins of the H-M populations. The higher ratio of northern lineages observed in the Miao people suggests that they had more contacts with the northern East Asians. Our systematic study of H-M mtDNA diversity provides genetic evidence for the origin and migration of the H-M populations and the data for further investigation of the genetic structure of East Asians.

Supplemental Material

Accession numbers for sequences reported here are as follows: AY546723–AY547259. Individual data of coding region variants are available as Supplemental Material online.

Acknowledgments

We thank all of the DNA donors for making this work possible. This work is partially supported by NSFC (39993420). L.J. is also supported by the National Science Foundation Under NSF grant BCS-0213857, and L.J. and R.C. are supported by the National Institutes of Health under NIH grant GM 41399.

Literature Cited

Andrews, R. M., I. Kubacka, P. F. Chinnery, R. N. Lightowlers, D. M. Turnbull, and N. Howell. 1999. Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA. Nat. Genet. **23**:147.

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

Bandelt, H. J., P. Lahermo, M. Richards, and V. Macaulay. 2001. Detecting errors in mtDNA data by phylogenetic analysis. Int. J. Legal Med. **115**:64–69.

Cang, M. 1997. Study on the migration culture of the ethnic groups in Yunan. Yunnan Nationality Press, Kunming (in Chinese).

Cavalli-Sforza, L. L., and M. W. Feldman. 2003. The application of molecular genetic approaches to the study of human evolution. Nat. Genet. **33**:266–275.

Cavalli-Sforza, L. L., P. Menozzi, and A. Piazza. 1994. The history and geography of human genes. Princeton University Press, Princeton, N.J.

Ding, Y-C., S. Wooding, H. Harpending et al. (11 co-authors). 2000. Population structure and history in East Asia. Proc. Natl. Acad. Sci. USA **97**:14003–14006.

Excoffier, L., P. E. Smouse, and J. M. Quattro. 1992. Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. Genetics **131**:479–491.

Fei, X. T. 1999. The pattern of diversity in unity of the Chinese nation. Central University for Nationalities Press, Beijing (in Chinese).

Forster, P., R. Harding, A. Torroni, and H-J. Bandelt. 1996. Origin and evolution of native American mtDNA variation: a reappraisal. Am. J. Hum. Genet. **59**:935–945.

Horai, S., K. Murayama, K. Hayasaka, S. Matsubayashi, Y. Hattori, G. Fucharoen, S. Harihara, K. S. Park, K. Omoto, and I. H. Pan. 1996. mtDNA polymorphism in East Asian Populations, with special reference to the peopling of Japan. Am. J. Hum. Genet. **59**:579–590.

Jin, L., and B. Su. 2000. Natives or immigrants: modern human origin in East Asia. Nat. Rev. Genet. **1**:126–133.

Karafet, T., L. Xu, R. Du, W. Wang, S. Feng, R. S. Wells, A. J. Redd, S. L. Zegura, and M. F. Hammer. 2001. Paternal population history of East Asia: sources, patterns, and microevolutionary process. Am. J. Hum. Genet. **69**:615–628.

Kivisild, T., H-V. Tolk, J. Parik, Y. Wang, S. S. Papiha, H-J. Bandelt, and R. Villems. 2002. The emerging limbs and twigs of the East Asian mtDNA tree. Mol. Biol. Evol. **19**:1737–1751.

Kong, Q. P., Y. G. Yao, C. Sun, H-J. Bandelt, C. L. Zhu, and Y. P. Zhang. 2003*a*. Phylogeny of East Asian mitochondrial DNA lineages inferred from complete sequences. Am. J. Hum. Genet. **73**:671–676.

Kong, Q. P., Y. G. Yao, M. Liu, S. P. Shen, C. Chen, C. L. Zhu, M. G. Palanichamy, and Y. P. Zhang. 2003*b*. Mitochondrial DNA sequence polymorphisms of five ethnic populations from northern China. Hum. Genet. **113**:391–405.

Kumar, S., K. Tamura, I. B. Jakobsen, and M. Nei. 2001. MEGA2: molecular evolutionary genetics analysis software. Bioinformatics **17**:1244–1245.

Lee, S., C. Shin, K. Kim, Y. Lee, and J. Lee. 1997. Sequence variation of mitochondrial DNA control region in Koreans. Forensic Sci. Int. **87**:99–116.

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

Oota, H., T. Kitano, F. Jin, I. Yuasa, L. Wang, S. Ueda, N. Saitou, and M. Stoneking. 2002. Extreme mtDNA homogeneity in continental Asian populations. Am. J. Phys. Anthropol. **118**:46–53.

Saillard, J., P. Forster, N. Lynnerup, H-J. Bandelt, and S. Nørby. 2000. mtDNA variation among Greenland Eskimos: the edge of the Beringian expansion. Am. J. Hum. Genet. **67**:718–726.

Schneider, S., J-M. Kueffer, D. Roessli, and L. Excoffier. 2000. Arlequin: v. 2.000A software for population genetic analysis. Genetics and Biometry Laboratory, University of Geneva, Geneva, Switzerland.

Snall, N., M. L. Savontaus, S. Kares, M. S. Lee, E. K. Cho, J. O. Rinne, and K. Huoponen. 2002. A rare mitochondrial DNA haplotype observed in Koreans. Hum. Biol. **74**:253–262.

Su, B., J. Xiao, P. Underhill et al. (21 co-authors). 1999. Y-chromosome evidence for a northward migration of modern humans into eastern Asia during the last ice age. Am. J. Hum. Genet. **65**:1718–1724.

Su, B., C. Xiao, R. Deka, R et al. (11 co-authors). 2000. Y chromosome haplotypes reveal prehistorical migrations to the Himalayas. Hum. Genet. **107**:582–590.

Tajima, A., C. S. Sun, I. H. Pan, T. Ishida, N. Saitou, and S. Horai. 2003. Mitochondrial DNA polymorphisms in nine aboriginal groups of Taiwan: implications for the population history of aboriginal Taiwanese. Hum. Genet. **113**:24–33.

Wallace, D. C., M. D. Brown, and M. T. Lott. 1999. Mitochondrial DNA variation in human evolution and disease. Gene **238**:211–230.

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

Wen, B., X. Xie, S. Gao et al. (13 co-authors). 2004*a*. Analyses of genetic structure of Tibeto-Burman populations revealed a gender-biased admixture in southern Tibeto-Burmans. Am. J. Hum. Genet. **74**:856–865.

Wen, B., H. Li, D. R. Lu et al. (17 co-authors). 2004*b*. Genetic evidence supports Demic diffusion of Han culture. Nature **431**:302–305.

Xiao, C. J., R. F. Du, L. L. Cavalli-Sforza, and E. Minch. 2000. Principal component analysis of gene frequencies of Chinese populations. Sci. China (Ser C) **43**:472–481.

Yao, Y. G., W. S. Watkins, and Y. P. Zhang. 2000. Evolutionary history of the mtDNA 9-bp deletion in Chinese populations and its relevance to the peopling of East and Southeast Asia. Hum. Genet. **107**:504–512.

Yao, Y. G., Q. P. Kong, H-J. Bandelt, T. Kivisild. and Y. P. Zhuang. 2002*a*. Phylogeographic differentiation of mitochondrial DNA in Han Chinese. Am. J. Hum. Genet. **70**: 635–651.

Yao, Y. G., L. Nie, H. Harpending, Y. X. Fu, Z. G. Yuan, and Y. P. Zhang. 2002*b*. Genetic relationship of Chinese ethnic populations revealed by mtDNA sequence diversity. Am. J. Phys. Anthropol. **118**:63–76.

Yao, Y. G., and Y. P. Zhang. 2002. Phylogeographic analysis of mtDNA variation in four ethnic populations from Yunnan Province: new data and a reappraisal. J. Hum. Genet. **47**: 311–318.

Yao, Y. G., Q. P. Kong, X. Y. Man, H. J. Bandelt, and Y. P. Zhang. 2003. Reconstructing the evolutionary history of China: a caveat about inferences drawn from ancient DNA. Mol. Biol. Evol. **20**:214–219.

Lisa Matisoo-Smith, Associate Editor

Accepted November 15, 2004