J Hum Genet (2002) 47:311–318 

ORIGINAL ARTICLE

**Yong-Gang Yao · Ya-Ping Zhang**

# Phylogeographic analysis of mtDNA variation in four ethnic populations from Yunnan Province: new data and a reappraisal

Received: February 8, 2002 / Accepted: March 7, 2002

**Abstract** Two sets of mitochondrial DNA (mtDNA) hypervariable segment I (HVS-I) data from four ethnic populations (Tibetan, Va, Dai, and Lahu) from Yunnan Province, China, were analyzed here by using phylogeographic methods. The results suggest that more attention should be paid to sampling methodology when addressing the genetic relationship and affinity among ethnic populations. Comparison of related data from different labs may serve as a check for the credibility of the data and will help discern the origin of the ethnic populations. Generally, Tibetan populations have more north-prevalent haplogroups (clades of the mtDNA phylogeny), while Dai and Lahu populations have high frequencies of south-prevalent haplogroups. The Vas, although autochthonous according to historical records, show signs of gene admixtures from northern and southern populations, for they harbor high frequencies of the south-prevalent haplogroup F and the north-prevalent haplogroup D as well as other northern mtDNA lineages such as M9 and G2a. The consanguineous marriage customs of the Lahu, together with possible genetic drift during this group's historical migration, left a conspicuous genetic imprint on its current gene pool.

**Key words** mtDNA · Phylogeography · Chinese ethnic population · Genetic structure · Ethnogenesis

## Introduction

In the past two decades, mitochondrial DNA (mtDNA) variation analyses have been extensively used to study the provenance and migration patterns of human populations (for reviews, see Wallace et al. 1999; Yao and Zhang 2000) and to estimate past population size and dates of expansion events (Harpending et al. 1998), as well as to study the phylogeographic differentiation and demographic history of single or multiple ethnic populations (Yao et al. 2002a,b). Usually, only a limited number of samples were analyzed in each population. On the other hand, due to possible competition and for other reasons, some populations have been sampled more than once and similar analyses have been performed in different labs. Leaving aside the scientific spirit of confirmation and checking, we ask ourselves whether such duplication is justifiable.

Recently, one data set of mtDNA hypervariable segment (HVS-I) data from four ethnic populations (Dai, Tibetan, Va, Lahu) from Yunnan Province, China, was published in two papers by Qian et al. (2001a,b), and in our recent work (Yao et al. 2002a), we also sequenced samples from Dai and Tibetan populations. It is worth comparing the two data sets now, to answer the aforementioned question. Moreover, the data analyses of both Qian et al. (2001a, b) and Yao et al. (2002a) were not sufficient: Qian and his colleagues mainly performed a neighbor-joining (NJ) analysis of the haplotypes, a method that has many problems, as we have recently described (Yao et al. 2002b) and which we will discuss in more detail here one; whereas in our paper (Yao et al. 2002a), we mainly used mismatch analyses (Rogers and Harpending 1992), neutrality tests (Tajima 1989; Fu 1997) and a principle component (PC) analysis of the populations, and we did not include a detailed analysis at the lineage level. Thus, some useful information was not uncovered. Moreover, these population genetic methods are sensitive to sampling methodology: small sample sizes may result in multimodal mismatch distributions and no deviations in the neutrality tests (Yao et al. 2002a), and thus mimic a constant population size in the past (Tajima 1989; Rogers and Harpending 1992; Fu 1997). By using what has become known as the "phylogeographic" approach, which has been well described in recent studies (Macaulay et al. 1999; Torroni et al. 2001; Yao et al. 2002b), we here reanalyze the two data sets from these same ethnic populations,

Y.-G. Yao · Y.-P. Zhang (✉)
Kunming Institute of Zoology, Chinese Academy of Sciences, Kunming, Yunnan 650223, China
Tel. +86-871-5195430; Fax +86-871-5195430
e-mail: zhangyp@public.km.yn.cn

Y.-G. Yao
Graduate School of the Chinese Academy of Sciences, Beijing, China

**Table 1.** Data sources and related information

| Population[a] | No. | Sample location | Reference | Accession numbers |
|---|---|---|---|---|
| Dai-1 (D-1) | 21 | Xishuangbanna, Yunnan | Qian et al. 2001a | — |
| Tibetan-1 (T-1) | 24 | Diqing, Yunnan | Qian et al. 2001a | — |
| Va-1 (V-1) | 22 | Simao, Yunnan | Qian et al. 2001a | — |
| Lahu-1 (L-1) | 32 | Simao, Yunnan | Qian et al. 2001a | — |
| Dai-2 (D-2) | 41 | Xishuangbanna, Yunnan | Yao et al. 2002a; current study | AF392131–392168; AF478680–478682 |
| Tibetan-2 (T-2) | 40 | Deqian, Diqing, Yunnan; Qinghai | Yao et al. 2002a | AF392245–392284 |
| Va-2 (V-2) | 36 | Ximeng and Gengma, Yunnan | Current study | AF478645–478679 |
| Lahu-2 (L-2) | 35 | Lancang, Yunnan | Current study | AF478609–478644 |

[a] The abbreviation for each sample population is in parentheses

with special emphases on (1) a comparison of the two data sets and (2) the maternal genetic structure and genetic relationships among the four populations.

## Material and methods

### Sampling

Since the Va and Lahu population samples were not included in our previous work (Yao et al. 2002a), we collected data from 38 Vas from Ximeng and Gengma, Yunnan Province, and from 35 Lahus from Lancang, Yunnan Province, to compare with the data reported by Qian et al. (2001a). Additionally, samples from three Dai from Xisuanbanna, Yunnan Province, were also sequenced here. All the individuals gave informed consent. Further sample information of the data considered here is listed in Table 1.

### HVS-I amplification, sequencing, and additional polymorphism typing

The HVS-I segments were amplified and sequenced as described in our previous studies (Yao et al. 2000a; 2002a). Then, the sequences were edited and aligned by using Dnastar software (DNASTAR, Madison, WI, USA) and compared with the revised Cambridge reference sequence (CRS, Andrews et al. 1999). The length polymorphisms of the A and C stretches in region 16180–16188 (triggered by the 16189 T to C substitution; nucleotide positions are relative to CRS) were disregarded in the analyses. Those individuals who had not yet been screened for the mtDNA 9-bp deletion in the COII/$tRNA^{Lys}$ intergenic region (Yao et al., 2000b, 2001) were typed for the deletion as described in those studies. In addition, five individuals (four of them bearing the HVS-I motif of haplogroup M9 and one without the motif) were also typed for the site 3394 mutation by using the same primers and conditions as described by Yao et al. (2002b).

### Data analyses

We used the same strategy as before (Yao et al. 2002b) to predict the potential haplogroup status of the previously reported mtDNAs (Qian et al. 2001a; Yao et al. 2002a) as well as that of the new data considered here: we classified the mtDNAs into haplogroups according to the HVS-I motifs and (near)matching with the 327 Han mtDNAs (Yao et al. 2002b; Kivisild et al. 2002). When naming haplogroups, we followed the spirit and process described in Yao et al. (2002b) and Kivisild et al. (2002). Since the characteristic motif of haplogroup M10 (Yao et al. 2002b) is based on a recurrent site (16311) in HVS-I (Hasegawa et al. 1993) and site 10646 was unavailable in the present data, we do not highlight M10 here. All the potential M10 types were just regarded as M* types.

We computed the haplogroup frequencies in the data set of Qian and his colleagues (2001a) and in our own data (Yao et al. 2002a, current study), to compare these two data sets from different labs and to determine whether there were any differences in the genetic components of the four ethnic populations. Then, haplogroup frequency profiles were subjected to principal component (PC) analyses to obtain a rough image of the maternal genetic relationship among the samples.

## Results and discussion

### Haplogroup identification

Table 2 displays the two data sets together with haplogroup classifications. In all, 25 (sub)haplogroups can be identified in the data sets, as shown in Fig. 1. One Va individual could not be further classified into the specific haplogroups of R that have been identified (R* in Table 2; Yao et al. 2002b; Kivisild et al. 2002). Two Dais had no mutation in the HVS-I sequence compared with the CRS (Andrews et al. 1999), but matched two Han individuals (one belonging to haplogroup F2, and the other to M7) from Guangzhou, Guangdong Province (Kivisild et al. 2002). Thus, they might be of the F2 or M7 type. Six individuals (2 Dais, 4 Lahus) bore the motif 16304-16309-16390 that occurred in one Han individual from Xinjiang, China (Yao et al. 2002b), which we named haplogroup R10, although no coding-region information or representative complete sequences are available for this haplogroup. The R10 motif also presented in two Taiwanese Han individuals (Tsai et al. 2001), and in two Zhuangs from Guangxi (Yuan et al. 2001; Yao et al. 2002a),

**Table 2.** Sequence variations in the two data sets

| Sample name | HVS-I variation (16,000+) | Haplogroup | Qian et al. 2001a | | | | Yao et al. 2002a; current study | | | | 9 bp | 3391e |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | D-1 | L-1 | T-1 | V-1 | D-2 | L-2 | T-2 | V-2 | | |
| 1 | 189 213 217 261 292 (519) | B4a | | 5 | | | | | | | 1 | |
| d29 | 189 213 217 261 292 362 | B4a | | | | | 1 | | | | 1 | |
| 2 | 179 189 213 217 261 292 362 (519) | B4a | 1 | | | | | | | | 1 | |
| 3 | 187 189 214 216 217 261 (519) | B4a | 1 | | | | | | | | 1 | |
| dai5 | 085G 093 188 189 214 217 261 287 | B4a | | | | | 1 | | | | 1 | |
| d43 | 129 189 217 261 | B4a | | | | | 1 | | | | 1 | |
| dai4 | 147 184A 189 217 235 | B4 | | | | | 1 | | | | 1 | |
| 6 | 147 184A 189 217 223 235 (519) | B4 | 1 | | | | | | | | 1 | |
| lahu71 | 189 217 223 (519) | B4 | | 3 | | | | 1 | | | 1 | |
| z63 | 189 217 256 311 362 | B4 | | | | | | | 1 | | 1 | |
| 5 | 136 140 189 217 249 274 280 291 335 (519) | B4 | | | | 1 | | | | | 1 | |
| d42 | 140 189 223 266A | B5a | | | | | 1 | | | | 1 | |
| dai83 | 140 189 266A | B5a | | | | | 3 | | | 1 | 1 | |
| 4 | 111 140 189 234 243 463 (519) | B5b | 1 | | | | | | | | 1 | |
| dai38 | 108 189 362 | B | | | | | 1 | | | | 1 | |
| dai52 | 129 189 352 355 | B | | | | | 1 | | | | 1 | |
| dai70 | 189 292 | B | | | | | 2 | | | | 1 | |
| 8 | 145 189 | R* | | | | 1 | | | | | 2 | |
| 21 | 172 213 304 (519) | F1a | 1 | | | | | | | | 2 | |
| 22 | 172 304 465 (519) | F1a | 1 | | | | | | | | 2 | |
| 23 | 172 304 (519) | F1a | 1 | | | | | | | | 2 | |
| d40 | 129 172 304 (519) | F1a | | 5 | | 1 | 1 | 13 | | 6 | 2 | |
| 24 | 129 172 189 304 (519) | F1a | 1 | | | | | | | | 2 | |
| Va82 | 111 129 172 304 | F1a | | | | | | | | 1 | 2 | |
| 20 | 111 129 304 (519) | F1a | | | 1 | | | | | | 2 | |
| d41 | 086 111 129 172 189 304 | F1a | | | | | 1 | | | | 2 | |
| Va66 | 066 129 172 304 | F1a | | | | | | | | 2 | 2 | |
| Va70 | 129 172 187 304 | F1a | | | | | | | | 3 | 2 | |
| lahu61 | 129 172 218A 304 | F1a | | | | | | 2 | | | 2 | – |
| 26 | 129 172 241 304 (519) | F1a | | | | 2 | | | | | 2 | |
| 27 | 129 162 172 293C 304 (519) | F1a | | | | 2 | | | | | 2 | |
| Va87 | 108 129 162 172 304 | F1a | | | | | | 3 | | 2 | 2 | |
| Va95 | 108 129 162 172 295 304 | F1a | | | | | | | | 1 | 2 | |
| 28 | 108 129 162 172 304 311 (519) | F1a | 1 | | | | | | | | 2 | |
| 29 | 108 129 162 172 262 304 (519) | F1a | | 5 | | 1 | | | | | 2 | |
| d47 | 108 129 162 172 189 304 | F1a | | | | | 1 | | | | 2 | |
| d72 | 108 124 129 162 172 304 | F1a | | | | | 1 | | | | 2 | |
| Va14 | 129 162 172 304 311 362 | F1a | | | | | | | | 1 | 2 | |
| d53 | 129 162 172 304 399 | F1a | | | | | 1 | | | | 2 | |
| 13 | 189 192 288 304 352 390 (519) | F1b | | | | 1 | | | | | 2 | |
| 12 | 126 189 304 318 (519) | F1b | | | 1 | | | | | | 2 | |
| d9 | 145 192 223 291 304 | F2a | | | | | 1 | | | | 2 | |
| d51 | 092A 170T 189 291 304 | F2a | | | | | 1 | | | | 2 | |
| lahu63 | 086 167 203 304 318 | F2 | | | | | | 8 | | | 2 | |
| lahu95 | 086 167 203 304 (519) | F2 | | 1 | | | | 1 | | | 2 | |
| d100 | 304 362 | F | | | | | 1 | | | | 2 | |
| d91[a] | CRS | F2 or M7 | | | | | 2 | | | | 2 | |
| 17 | 124 148 304 390 (519) | R10 | | 1 | | | | | | | 2 | |
| 16 | 093 209 304 309 390 | R10 | | 1 | | | | | | | 2 | |
| 18 | 093 148 209 304 309 390 | R10 | | 1 | | | | | | | 2 | |
| d99 | 192 239 304 309 390 | R10 | | | | | 1 | | | | 2 | |
| 15 | 192 304 309 390 (519) | R10 | | 1 | | | | | | | 2 | |
| 14 | 051 192 304 309 390 (519) | R10 | 1 | | | | | | | | 2 | |
| d7 | 126 189 223 298 311 327 | C | | | | | 1 | | | | 2 | |
| d3 | 189 223 298 319 327 | C | | | | | 1 | | | | 2 | |
| d109 | 189 223 298 327 | C | | | | | 1 | | | | 2 | |
| d10 | 223 298 327 | C | | | | | 1 | | | | 2 | |
| 30 | 223 298 327 355 (519) | C | 1 | | | | | | | | 2 | |
| Dai117 | 129 148 223 242 298 311 319 327 | C | | | | | 1 | | | | 2 | |
| 31 | 185 189 223 260 298 302 | Z | | | 1 | | | | | | 2 | |
| 33 | 185 223 260 298 355 | Z | | | 2 | | | | | | 2 | |
| 32 | 185 223 260 298 | Z | | | 3 | | | | | | 2 | |
| lahu85 | 260 298 355 362 | Z | | | | | | 1 | | | 2 | |
| z5 | 215 223 274 399 | M* | | | | | | | 3 | | 2 | |
| z8 | 215 223 274 356 | M* | | | | | | | 1 | | 2 | |
| 52 | 215 223 274 (16T region 1–41) | M* | | | 1 | | | | | | 2 | |
| 51 | 173 209 223 265C 267A | M* | | | | 1 | | | | | 2 | |

**Table 2.** Continued

| Sample name | HVS-I variation (16,000+) | Haplogroup | Qian et al. 2001a | | | | Yao et al. 2002a; current study | | | | 9bp | 3391e |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | D-1 | L-1 | T-1 | V-1 | D-2 | L-2 | T-2 | V-2 | | |
| Va69 | 129 209 223 272 | M* | | | | | | | | 3 | 2 | |
| 41 | 129 140 203 223 271 | M* | | 2 | | | | | | | 2 | |
| d1 | 075 085G 223 293T | M* | | | | | 1 | | | | 2 | |
| 10 | 093 145 148 188 189 223 381 | M* | | | 1 | | | | | | 2 | |
| 9 | 145 148 188 189 223 381 (519) | M* | | | 1 | | | | | | 2 | |
| 11 | 145 148 188 189 223 381 | M* | | | 2 | | | | | | 2 | |
| z76 | 145 168 188 223 257 311 | M* | | | | | | | 1 | | 2 | |
| 62 | 223 234 290 325 362 | M* | 1 | | | | | | | | 2 | |
| 43 | 129 172 223 290 325 354 (519) | M* | | 4 | | | | | | | 2 | |
| Va90 | 093 129 223 234 290 311 | M* | | | | | | | | 1 | 2 | |
| Va9 | 111 129 183 189 223 311 468 | M* | | | | | | | | 1 | 2 | |
| 36 | 093 223 249 266 301 311 (519) | M* | | | | 1 | | | | | 2 | |
| 37 | 223 249 266 301 311 (519) | M* | | | | 1 | | | | | 2 | |
| z7 | 129 189 223 249 311 | M* | | | | | | | 1 | | 2 | |
| z36 | 223 311 | M* | | | | | | | 5 | | 2 | |
| lahu73 | 223 269 271 311 | M* | | | | | | 1 | | | 2 | |
| 38 | 182 193 223 (519) | M* | | | | 4 | | | | 1 | 2 | |
| Va76 | 182 193 223 297 | M* | | | | | | | | 1 | 2 | |
| d35 | 086 223 297 | M7b | | | | | 1 | | | | 2 | |
| d48 | 129 189 223 297 | M7b | | | | | 1 | | | | 2 | |
| d6 | 129 189 223 248 297 300 | M7b | | | | | 1 | | | | 2 | |
| 40 | 129 189 223 297 468 (519) | M7b | 1 | | | | | | | | 2 | |
| d84 | 129 192 223 297 | M7b1 | | | | | 1 | | | | 2 | |
| d44 | 129 192 223 297 357 | M7b1 | | | | | 1 | | | | 2 | |
| 39 | 129 192 223 297 305 | M7b1 | 1 | | | | | | | | 2 | |
| 34 | 176 223 295 (519) | M7c | 1 | | | | | | | | 2 | |
| 35 | 223 260 295 (519) | M7c | | | 1 | | | | | | 2 | |
| z4 | 169 223 260 295 | M7c | | | | | | | 1 | | 2 | |
| z61 | 093A 145 223 234 316 354 | M9 | | | | | | | 1 | | 2 | + |
| z70 | 223 234 316 362 | M9 | | | | | | | 1 | | 2 | + |
| z6 | 158 223 234 362 (519) | M9 | | | 1 | | | | 1 | | 2 | + |
| Va67 | 086 158 223 234 362 | M9 | | | | | | | | 3 | 2 | + |
| 65 | 158 223 234 287 362 (519) | M9 | | | 1 | | | | | | 2 | |
| 64 | 093 158 223 234 362 (519) | M9 | | | 1 | | | | | | 2 | |
| 63 | 223 234 266 316 362 | M9 | | | 1 | | | | | | 2 | |
| d50 | 184 189 223 319 470 471 473 | M8a | | | | | 1 | | | | 2 | |
| 59 | 223 224 227 278 362 | G2a | | | | 1 | | | | | 2 | |
| 56 | 129 164 189 223 266 362 | D5 | | | | 1 | | | | | 2 | |
| z55 | 092 164 172 189 223 243 266 362 | D5 | | | | | | | 1 | | 2 | |
| 57 | 164 172 189 223 266 362 | D5 | | 1 | | | | | | | 2 | |
| 55 | 131 185 189 223 232A 319 320 362 | D5 | 1 | | | | | | | | 2 | |
| d46 | 189 223 362 | D5 | | | | | 2 | | | | 2 | |
| z54 | 111 172 189 223 362 | D5 | | | | | | | 1 | | 2 | |
| d32 | 086 092 362 | D | | | | | 1 | | | | 2 | |
| Va94 | 223 274 362 | D | | | | | | | | 1 | 2 | |
| z9 | 169 223 274 362 | D | | | | | | | 1 | | 2 | |
| Va11 | 223 274 311 362 | D | | | | | | | | 1 | 2 | |
| 53 | 223 274 356 362 (519) | D | | | | 2 | | | | | 2 | |
| 54 | 086 223 274 291 362 (519) | D | | 2 | | | | 5 | | | 2 | |
| 58 | 223 245 362 | D | 1 | | | | | | | | 2 | |
| Va84 | 223 245 311 362 | D | | | | | | | | 1 | 2 | |
| d12 | 223 256 311 362 | D | | | | | 1 | | | | 2 | |
| 42 | 129 223 256 362 (519) | D | 1 | | | | | | | | 2 | |
| 61 | 093 223 294 362 | D | | | | 1 | | | | | 2 | |
| z90 | 093 223 362 | D | | | | | | | 1 | | 2 | |
| z29 | 093 223 319 362 | D | | | | | | | 7 | | 2 | |
| z45 | 223 319 362 | D | | | | | | | 5 | | 2 | |
| z39 | 223 241 269 362 | D | | | | | | | 4 | | 2 | |
| Va10 | 111 223 362 | D | | | | | | | | 1 | 2 | |
| Va78 | 111 172 193 223 362 | D | | | | | | | | 1 | 2 | |
| Va83 | 172 193 223 362 | D | | | | | | | | 2 | 2 | |
| z85 | 223 362 | D | | | | | | | 2 | 1 | 2 | |
| Va15 | 192 223 316 362 | D | | | | | | | | 1 | 2 | |
| 60 | 093 192 223 316 362 | D | 1 | | | | | | | | 2 | |
| 44 | 223 257A 261 292 319 (519) | N9a | 1 | | | | | | | | 2 | |
| 45 | 111 129 223 257A 261 319 | N9a | 1 | | | | | | | | 2 | |
| 46 | 223 234 290 293C 319 (519 527) | A | | | 3 | | | | | | 2 | |

**Table 2.** Continued

| Sample name | HVS-I variation (16,000+) | Haplogroup | Qian et al. 2001a | | | | Yao et al. 2002a; current study | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | D-1 | L-1 | T-1 | V-1 | D-2 | L-2 | T-2 | V-2 | 9bp | 3391e |
| 47 | 223 290 294 295 319 362 (519) | A | | | 1 | | | | | | 2 | |
| z28 | 192 223 290 295 319 362 | A | | | | | | | 2 | | 2 | |
| 48 | 223 278 290 319 362 | A | | | | 1 | | | | | 2 | |
| d45 | 223 290 319 362 | A | | | | | 1 | | | | 2 | |
| d37 | 093 129 145 223 290 319 362 | A | | | | | 1 | | | | 2 | |
| 49 | 170 223 274 290 319 362 (527) | A | | | 1 | | | | | | 2 | |
| 50 | 223 274 290 319 362 (527) | A | | | 1 | | | | | | 2 | |

Haplogroup status of each mtDNA was predicted according to the strategy described in Yao et al. (2002b). Those haplotypes with many ambiguities in haplogroup classification (including the potential M10 types) but having the 16223 mutation were regarded as M* types. Those mutations present in the fragments [from site 16048 to site 41 relative to the numbering of the Cambridge reference sequence (CRS)] sequenced by Qian et al. (2001a), but not in the fragments (from site 16001 to 16497) that we sequenced are in parentheses. The restriction enzyme *Hae*III that was used for site 3394 typing in haplogroup M9 is designated by "e." The minus and plus signs denote the absence and presence, respectively, of the restriction site. The length polymorphisms of the A and C stretches in region 16180–16188 (triggered by the 16189 T to C substitution) are not included in the table

[a] These two mtDNAs had no mutation compared with the CRS (Andrews et al. 1999). Although they might be F2, M7, or another type, we regarded them as F2 types for the haplogroup frequency computation

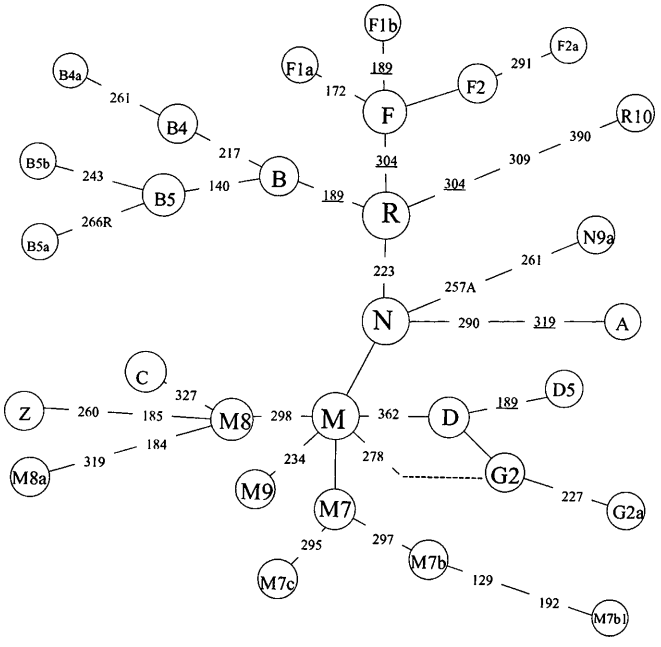


**Fig. 1.** Schematic profile of the mtDNA haplogroups in the four ethnic populations. The HVS-I–specific mutations (motif) of each haplogroup [nucleotide positions are relative to the Cambridge reference sequence (Andrews et al. 1999), but minus 16,000] are indicated on the *links*, but the order of the mutations on a *link* is arbitrary. Transversions at sites 16257 and 16266 are specified by adding the suffixes *A* and *R*, respectively; recurrent mutations in the figure are *underlined*

but it is very rare in the reported data from other northern populations or populations of northern origin (Schurr et al. 1999; Yao et al. 2000a, 2002a, b). Therefore, haplogroup R10 tends to be restricted to southern populations. Twenty-two haplotypes had many uncertainties when we considered their HVS-I variations and were classified as M* types. In these M* types, several clades with specific motifs could be discerned: 16223-16249-16311 was found in two Vas and one Tibetan; 16215-16223-16274 and 16145-16188-16223 all appear specific to Tibetan individuals; 16223-(16234)-16290 occurred in one Dai, four Lahus, and one Va; 16182-16193-16223 was present in six Vas. The detailed phylogenetic positions of these little clades (subhaplogroups of M) await more information from coding regions or complete sequences.

The haplogroup classification of the mtDNAs contradicts the results of the phylogenetic analysis performed by Qian et al. (2001a) by using the NJ method: in the eight clusters seen in their Figure 2, F types were divided into clusters 3 and 4 and D types were separated into clusters 7 and 8. The seventh cluster is very artificial from a phylogenetic point of view, for it contains types (M7b, D, M*, N9a, and A) from both macro-haplogroups M and N. Cluster 8 is also paraphyletic, for it harbors M*, D, D5, G2a, and M9 types. Thus, the apparent clusters in the NJ tree without reference to mtDNA phylogeny should be taken with extreme caution (Yao et al. 2002b).

## Errors detected in the data sets

For many reasons, errors in the published HVS-I data were not infrequent, as was summarized by Bandelt et al. (2001). By constructing a network relationship of the haplotypes (Bandelt et al. 1995, 2000) and (near)matching them with published data sets, the potential errors in the data could be well distinguished. In the data of Qian et al. (2001a; their Figure 1), several mistakes caused by base shifts are evident: (1) one Z haplotype (31) has a base shift at site 16185, and the original 16186 mutation is thus a mistake; (2) two D5 haplotypes (56 and 57) have a base shift at site 16189 so that the 16188 mutations in these two types are mistakes; and (3) haplotype 57 also has base shifts at sites 16164 and 16172, and the 16165 and 16173 mutations are mistakes. Note that all these obvious errors were found in both papers of Qian et al. (2001a, b). Moreover, Qian et al. (2001a) incorrectly identified type 57 as from a Dai individual and type 58 as from a Lahu individual in their Figure 1.

**Table 3.** Haplogroup frequency profiles of the four ethnic populations

| | Qian et al. 2001a | | | | Yao et al. 2002a; current study | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Haplogroup | D-1 | L-1 | T-1 | V-1 | D-2 | L-2 | T-2 | V-2 | Dai (D) | Lahu (L) | Tibetan (T) | Va (V) |
| B4a | 9.5 | 15.6 | | | 7.1 | | | | 8.1 | 7.5 | | |
| B4* | 4.8 | 9.4 | | 4.5 | 2.4 | 2.9 | 2.5 | | 3.2 | 6.0 | 1.6 | 1.7 |
| B5a | | | | | 9.5 | | | 2.8 | 6.5 | | | 1.7 |
| B5b | 4.8 | | | | | | | | 1.6 | | | |
| B* | | | | | 9.5 | | | | 6.5 | | | |
| R* | | | | 4.5 | | | | | | | | 1.7 |
| F1a | 23.8 | 31.3 | 4.2 | 27.3 | 11.9 | 51.4 | | 44.4 | 16.1 | 41.8 | 1.6 | 37.9 |
| F1b | | | 4.2 | 4.5 | | | | | | | 1.6 | 1.7 |
| F2a | | | | | 4.8 | | | | 3.2 | | | |
| F2* | | 3.1 | | | 4.8 | 25.7 | | | 3.2 | 14.9 | | |
| F | | | | | 2.4 | | | | 1.6 | | | |
| R10 | 4.8 | 12.5 | | | 2.4 | | | | 3.2 | 6.0 | | |
| C | 4.8 | | | | 11.9 | | | | 9.7 | | | |
| Z | | | 25.0 | | | 2.9 | | | | 1.5 | 9.4 | |
| M* | 4.8 | 18.8 | 20.8 | 31.8 | 2.4 | 2.9 | 27.5 | 19.4 | 3.2 | 10.4 | 25.0 | 24.1 |
| M7b1 | 4.8 | | | | 4.8 | | | | 4.8 | | | |
| M7b* | 4.8 | | | | 7.1 | | | | 6.5 | | | |
| M7c | 4.8 | | 4.2 | | | | 2.5 | | 1.6 | | 3.1 | |
| M8a | | | | | 2.4 | | | | 1.6 | | | |
| M9 | | | 16.7 | | | | 7.5 | 8.3 | | | 10.9 | 5.2 |
| G2a | | | | 4.5 | | | | | | | | 1.7 |
| D5 | 4.8 | 3.1 | | 4.5 | 4.8 | | 5.0 | | 4.8 | 1.5 | 3.1 | 1.7 |
| D* | 14.3 | 6.3 | | 13.6 | 4.8 | 14.3 | 50.0 | 25.0 | 8.1 | 10.4 | 31.3 | 20.7 |
| N9a | 9.5 | | | | | | | | 3.2 | | | |
| A | | | 25.0 | 4.5 | 4.8 | | 5.0 | | 3.2 | | 12.5 | 1.7 |

Haplogroup frequencies in the pooled samples of each ethnic population are listed in the last four columns

## Haplogroup profiles and PC analysis

Table 3 presents the haplogroup frequency profiles of the two data sets and of the pooled samples from each ethnic population. Haplogroup D, which is absent in the Tibetan data of Qian et al. (2001a), was found in about 50% of individuals in our Tibetan data. Haplogroups Z and A, which are either absent or present with relatively low frequency in our Tibetan population, are present with considerably high frequencies (25%) in the Tibetan data of Qian et al. (2001a). The fact that most of our Tibetan subjects (32 individuals) were from an isolated village in the hilly region of Deqin County (Yao et al. 2002a) as well as the large Tibetan population (>4,600,000, 1990 census; Du and Yip 1993) could account for the differences between the two data sets. Haplogroup M9, which has high frequency in northern populations (Yao et al. 2002b), is found in both data sets of Tibetan mtDNAs, and is thus one of the major components of the Tibetan gene pool.

Although no statistically significant differences were found between the other three pairs of sample populations (Dai-1 vs Dai-2, Lahu-1 vs Lahu-2, and Va-1 vs Va-2), some features should still be mentioned: haplogroup N9a, which might have a southern origin and is present mainly in populations of ancient Pai-Yuai tribe origin (author's unpublished data), was found only in two Dai individuals by Qian et al. (2001a); haplogroup G2a, which is prevalent in northern or northwestern populations, was found in one Va by Qian et al. (2001a); the newly defined haplogroup, R10, presented at a high frequency (12.5%) in the Lahu sample of Qian et al. (2001a), but was not found in our Lahu sample. Because the sample sizes in the data of Qian and his colleagues were relatively small (with the exception of the Lahu, 32 individuals; the sample size was ≤24 for each of the other three populations), the differences observed here may be due either to insufficient sampling or potential regional differences and should be taken with caution.

The PC analysis of the haplogroup profiles of the two data sets (Table 3) presents a general view of the differences considered (Fig. 2). The relative locations of the four sample populations in the data of Qian et al. (2001a) and in our data (Yao et al. 2002a; current study) are roughly the same on the PC map, although the differences between the two Tibetan population samples and between the two Dai population samples are quite obvious and more pronounced. We also performed a PC analysis of the haplogroup profiles of these eight population samples and of the four pooled samples of each ethnic population, but the result was hardly changed (map not shown). Thus, the relative genetic relationships among the four ethnic populations could be demonstrated by both data sets, more or less.

## Genetic structure and demographic events

On the basis of the presence of certain lineages in each pooled sample of the four ethnic populations (Table 2) and the haplogroup profiles (Table 3), we may tentatively infer the respective gene pool of the four populations. The Tibetan population, which has ancestral affinity to the ancient nomadic tribes of northwest China (Du and Yip 1993), contains high frequencies (>60%) of haplogroups preva-

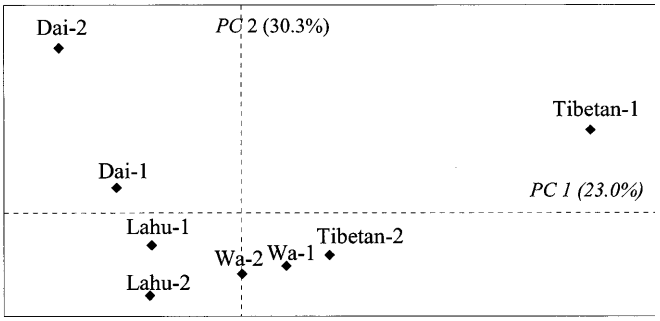


**Fig. 2.** Principal component (PC) map of the two data sets based on the mtDNA haplogroup profile shown in Table 3. The data from Qian et al. (2001a) and our own data (Yao et al. 2002a; current study) are marked by adding "*−1*" or "*−2*," respectively, after the *sample names*

lent in northwest, north, and central China, such as D, M9, A, and Z. Thus, it is a typically northern population. In the Dai, of ancient south Pai-Yuei tribe origin, the south-prevalent haplogroups, such as F, B, and M7b, have high frequencies and occupy more than 60% of this population's genetic pool. Haplogroup C, which is prevalent in northeast Asia (Schurr et al. 1999) and is one of the five Amerindian founder genes (Forster et al. 1996), presents with a relatively high frequency in the Dai (9.7%). The Vas are regarded as the earliest inhabitants of southwest Yunnan, and they speak a Mon-Khmer language that belongs to the south Asian language family (Du and Yip 1993). However, the maternal genetic structure of the Va bears the signatures of both southern and northern or northwestern populations: it contains relatively high frequencies of D lineages (>20%) as well as G2a and M9 lineages; on the other hand, it harbors haplogroup F1a at high frequency (37.9%). We suspect that the Va might have received genetic contributions from north or northwest China, possibly during the southwest migration of the ancient Di-Qiang tribe that occurred around 4000–5000 and 2000–2500 years ago (Du and Yip 1993; You 1994), but this suggestion is not supported by archaeological artifacts, language, or historical records.

The Lahus trace their origin to the ancient Di-Qiang tribe and were distributed in the Erhai area of Yunnan around the 8th century (Du and Yip 1993; You 1994). When the Nanzhao regime was flourishing in this region (A.D. 649–902) the ancestors of the Lahu (which were divided into three main ethnic branches: Lahu Xi, Lahu Pu, and Lahu Na) were compelled to move further southward, with the Lahu Xi and Pu people later dispersing in the area between Jindong, Pu'er, and Yuanjiang counties, and the Lahu Na people settling in Lanchang County and adjacent regions (Du and Yip 1993). Subsequent migrations of the Lahus were not on such a large scale (Du and Yip 1993). The genetic evidence provided here gives no direct support to the historical affinity of the Lahu with the ancient Di-Qiang tribe: the south-prevalent haplogroups B, F, and R10 occupy more than 60% of their gene pool, whereas in other populations of Di-Qiang origin, such as the Tibetan, Nu, and Lisu populations, more than 60% of lineages are of the D, A, C, and Z types (Yao et al. 2002a). However, the populations of Di-Qiang tribe origin might have undergone recurrent genetic drift during their southward migrations and subsequent settlement, which may account for the lower diversity observed in the sample populations (Yao et al. 2002a). As shown in Table 2, only 18 haplotypes were found in 67 Lahus, and one haplotype (d40) occurred in 18 individuals, while in the Dai sample population, 52 haplotypes were identified in 62 individuals. Moreover, the Lahus have a high frequency of consanguineous marriage (Du and Yip 1993). This nuptial custom would enhance the effect of genetic drift. Therefore, the possibility that a genetic imprint of the ancient Di-Qiang tribe in the Lahu was erased by a strong genetic drift effect can not be excluded. The following scenario for the Lahu is compatible with the results observed here: during their migrations and settlement, the original founder lineages of the Lahus might have been lost (at least in part) because of the instability and fragmentation of the population; at the same time, the Lahus might have assimilated genes from the local people, and the lineages that remained in their gene pool might have drifted further because of the consanguineous marriage custom.

The nucleotide sequence data reported here have been submitted to the GenBank database under the accession numbers AF478609–AF478682.

## Conclusion

The origin and development of ethnic population groups is a very complex process. Hypotheses that are based on different (limited) data sets or different analysis methods may contradict each other, (cf., the hot debates concerning the peopling of East Asia: Su et al. 1999; Ding et al. 2000; Karafet et al. 2001; Yao et al. 2002b), and this enlivens endless interest as well as serves as an impetus for the development of the field. Hence, similar studies on the same populations are necessary and may bring us to a better understanding about the past. The reanalyses of the two data sets suggests that both the studies of Qian et al. (2001a) and those of Yao et al. (2002a) might not be sufficient for fully reconstructing the real story. Sampling methodology is very important when addressing the genetic relationships and structures of ethnic populations, because potential regional differences resulting from different retained amounts of ancient genetic material as well as from the effects of later demographic events, such as genetic drift during migration and settlement and gene assimilation, might hinder our ability to recognize the whole story of the entire ethnic groups labeled Dai, Lahu, Tibetan, and Va. Therefore, the validity of results is restricted to the regional sample populations analyzed (Yao et al. 2002a). On the other hand, from this regional gene pool, a glimpse of the past is not impossible. When more data are available, we may finally get closer to the shadowy past lost in the long corridors of deep time.

**Acknowledgments** We thank YP Qian and colleagues for confirmation of the mistakes in their data. We also thank Q-P Kong for doublechecking the data presented here, and S-K Gou for technical assistance. This study was supported by grants from the Natural Science Foundation of Yunnan Province and the Natural Science Foundation of China (NSFC).

## References

Andrews RM, Kubacka I, Chinnery PF, Lightowlers RN, Turnbull DM, Howell N (1999) Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA. Nat Genet 23:147

Bandelt H-J, Forster P, Sykes BC, Richards MB (1995) Mitochondrial portraits of human populations using median networks. Genetics 141:743–753

Bandelt H-J, Macaulay V, Richards M (2000) Median networks: speedy construction and greedy reduction, one simulation, and two case studies from human mtDNA. Mol Phylogenet Evol 16:8–28

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

Ding Y-C, Wooding S, Harpending H, Chi H-C, Li H-P, Fu Y-X, Pang J-F, Yao Y-G, Xiang YJG, Moyzis R, Zhang Y-P (2000) Population structure and history in East Asia. Proc Natl Acad Sci U S A 97:14003–14006

Du R, Yip VF (1993) Ethnic groups in China. Science Press, Beijing

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

Fu Y-X (1997) Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection. Genetics 147:915–925

Harpending HC, Batzer MA, Gurven M, Jorde LB, Rogers AR, Sherry ST (1998) Genetic traces of ancient demography. Proc Natl Acad Sci U S A 95:1961–1967

Hasegawa M, Di Rienzo A, Kocher TD, Wilson AC (1993) Toward a more accurate time scale for the human mitochondrial DNA tree. J Mol Evol 37:347–354

Karafet T, Xu L, Du R, Wang W, Feng S, Wells RS, Redd AJ, Zegura SL, Hammer MF (2001) Paternal population history of East Asia: sources, patterns, and microevolutionary process. Am J Hum Genet 69:615–628

Kivisild T, Tolk H-V, Parik J, Wang Y, Papiha SS, Bandelt H-J, Villems R (2002) The emerging limbs and twigs of the East Asian mtDNA tree. Mol Biol Evol (in press)

Macaulay VA, Richards MB, Hickey E, Vega E, Cruciani F, Guida V, Scozzari R, Bonné-Tamir B, Sykes B, Torroni A (1999) The emerging tree of west Eurasian mtDNAs: a synthesis of control-region sequences and RFLPs. Am J Hum Genet 64:232–249

Qian YP, Chu Z-T, Dai Q, Wie C-D, Chu JY, Tajima A, Horai S (2001a) Mitochondrial DNA polymorphism in Yunnan nationalities in China. J Hum Genet 46:211–220

Qian YP, Chu Z-T, Dai Q, Wie C-D, Chu JY, Horai S (2001b) Genetic relationship among four Yunnan populations revealed by sequence of mtDNA D-loop. Acta Genet Sinica 28:291–300 (in Chinese)

Rogers AR, Harpending H (1992) Population growth makes waves in the distribution of pairwise genetic differences. Mol Biol Evol 9:552–569

Schurr TG, Sukernik RI, Starikovskaya YB, Wallace DC (1999) Mitochondrial DNA variation in Koryaks and Itel'men: population replacement in the Okhotsk Sea–Bering Sea region during the Neolithic. Am J Phys Anthropol 108:1–39

Su B, Xiao J, Underhill P, Deka R, Zhang W, Akey J, Huang W, Shen D, Lu D, Luo J, Chu J, Tan J, Shen P, Davis R, Cavalli-Sforza L, Chakraborty R, Xiong M, Du R, Oefner P, Chen Z, Jin L (1999) Y-Chromosome evidence for a northward migration of modern humans into eastern Asia during the last ice age. Am J Hum Genet 65:1718–1724

Tajima F (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123:585–595

Torroni A, Bandelt H-J, Macaulay V, Richards M, Cruciani F, Rengo C, Martinez-Cabrera V, Villems R, Kivisild T, Metspalu E, Parik J, Tolk HV, Tambets K, Forster P, Karger B, Francalacci P, Rudan P, Janicijevic B, Rickards O, Savontaus ML, Huoponen K, Laitinen V, Koivumaki S, Sykes B, Hickey E, Novelletto A, Moral P, Sellitto D, Coppa A, Al-Zaheri N, Santachiara-Benerecetti AS, Semino O, Scozzari R (2001) A signal, from human mtDNA, of postglacial recolonization in Europe. Am J Hum Genet 69:844–852

Tsai LC, Lin CY, Lee JC, Chang JG, Linacre A, Goodwin W (2001) Sequence polymorphism of mitochondrial D-loop DNA in the Taiwanese Han population. Forensic Sci Int 119:239–247

Wallace DC, Brown MD, Lott MT (1999) Mitochondrial DNA variation in human evolution and disease. Gene 238:211–230

Yao Y-G, Zhang Y-P (2000) Mitochondrial DNA and human evolution. Zool Res 21:392–406 (in Chinese)

Yao Y-G, Lü X-M, Luo H-R, Li W-H, Zhang Y-P (2000a) Gene admixture in the silk road of China — evidence from mtDNA and melanocortin 1 receptor polymorphism. Genes Genet Syst 75:173–178

Yao Y-G, Watkins WS, Zhang Y-P (2000b) 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, Yuan Z-G, Zhou Z-D, Geng P-L, Li Q-W, Zhang Y-P (2001) Frequency of the mtDNA 9-bp deletion among Chinese ethnic groups. Prog Nat Sci 11:358–364

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

Yao Y-G, Kong Q-P, Bandelt H-J, Kivisild T, Zhang Y-P (2002b) Phylogeographic differentiation of mitochondrial DNA in Han Chinese. Am J Hum Genet 70:635–651

You Z (1994) Yunnan Minzu Shi (History of Yunnan nationalities). Yunnan University Press, Kunming, China (in Chinese)

Yuan Z-G, Yao Y-G, Ma Z-X, Pang Q-P, Lie Y-R, Ma J, Zhang Y-P (2001) Mitochondrial DNA sequence variations of Zhuang ethnic group in Guangxi. Acta Genet Sinica 28:95–102 (in Chinese)